DOCUMENT RESUME

ED 154 674 HE 009 945

AUTHOR	Duck, Gary A.
TITLE	Planned Variations Study. Volume V: Student Performance Measures for Evaluating Secondary and Postsecondary Intervention Programs.
INSTITUTION	System Development Corp., Santa Monica, Calif. Studies and Evaluation Dept.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
REPORT NO	TM-5857/005/00
PUB DATE	31 Jan 78
CONTRACT	300-76-0539
NOTE	106p.; For related documents see HE 009 941-946
EDRS PRICE	MF-$0.83 HC-$6.01 Plus Postage.
DESCRIPTORS	Achievement Tests; Affective Tests; *Compensatory Education; *Disadvantaged Youth; *Evaluation Methods; *Intervention; *Measurement Instruments; Measurement Techniques; Performance Criteria; Performance Tests; Post Secondary Education; *Program Evaluation; Secondary Education; Study Skills

ABSTRACT

System Development Corporation conducted an extensive project of conceptualization and planning of models for compensatory educational intervention for disadvantaged students in secondary and postsecondary education. The present report focuses on the procedures used to identify and evaluate as available student performance measures that can be considered as candidate instruments for use in evaluating these compensatory programs. The tests include achievement tests, study skills tests, and affective tests. (Author/MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

System Development Corporation

ED154674

PLANNED VARIATIONS STUDY

STUDENT PERFORMANCE MEASURES FOR EVALUATING SECONDARY AND POSTSECONDARY INTERVENTION PROGRAMS

VOLUME V

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

IHE 089 945

JANUARY 31, 1978

TM-5857/005/00

PLANNED VARIATIONS STUDY

VOLUME V: STUDENT PERFORMANCE MEASURES FOR EVALUATING SECONDARY AND POSTSECONDARY INTERVENTION PROGRAMS

Gary A. Duck

Studies and Evaluation Department
System Development Corporation

January 31, 1978

TM-5857/005/00

The work on which this publication is based was performed pursuant to Contract 300-76-0539, as amended, with the U.S. Office of Education, Department of Health, Education, and Welfare. Views or conclusions contained in this study should not be interpreted as representing the official opinion or policy of the Department of Health, Education, and Welfare.

CONTENTS

	Page
INTRODUCTION	1
Problem	2
Organization of Report	3
EVALUATION STUDY REQUIREMENTS	4
LANGUAGE ARTS, READING, AND MATHEMATICS ACHIEVEMENT TESTS	7
Screening Criteria and Scoring Procedures	7
Summary of Assessment Results	12
Recommendations	15
Assessing Writing Skills	17
STUDY SKILLS TESTS	19
Screening Criteria and Scoring Procedures	19
Summary of Assessment Results	22
Recommendation	26
STUDENT AFFECTIVE MEASURES	27
Screening Criteria and Scoring Procedures	27
Summary of Results and Recommendations	32
PROBLEMS ENCOUNTERED IN THE INSTRUMENTATION PROCESS	38
Content of Achievement Tests	38
Identifying Affective Measures	39
Locating New Affective Instruments	39
Examining Tests for Bias	40
Appendix A: Achievement Test Summaries	A-1
Appendix B: Studies Skills Test Summaries	B-1
Appendix C: Affective Test Summaries	C-1

INTRODUCTION

During the period of September 1, 1976, through January 30, 1978, System Development Corporation (SDC) conducted an extensive planning effort under a contract with the U. S. Office of Education (USOE). Called the Planned Variations Study, a major dimension of the project involved the conceptualization and planning of models for compensatory educational intervention at the secondary and postsecondary levels. The project also included the planning of a longitudinal evaluation of the models, incorporating the design of an evaluation and analytic procedures to be used, the methodology for executing an evaluation study in a variety of field settings, and the selection of appropriate instrumentation to determine the effects of models on participating students.

The genesis for the Planned Variations Study arose from earlier evaluations of two programs operated by the Bureau of Higher and Continuing Education (BHCE), the Upward Bound program for secondary students and the Special Services program for postsecondary students. Based on the results of these earlier evaluations, which had indicated that both Upward Bound and Special Services were achieving certain of their objectives but could benefit from the development of alternative models, the Office of Planning, Budgeting and Evaluation (OPBE) contracted with SDC to carry out the planning activities outlined above.

In the 17 months of the Planned Variations Study, SDC performed the following tasks, each of which is described in a separate report:

- Developed two intervention strategies for compensatory education projects to meet the unique needs of disadvantaged secondary and postsecondary students.
- Developed the design of a longitudinal study to evaluate the effects of models growing out of the two intervention strategies, along with analytic procedures to answer the issues to be addressed in such a study.
- Developed a field implementation plan for a longitudinal study, by which intervention models could be implemented in a variety of settings and data collected as called for in a longitudinal study design.

- Developed a set of recommendations for instruments to measure student outcomes in a longitudinal evaluation, based on an analysis of available measures in the cognitive and affective domains.
- Developed an index of components of existing projects and a system for regularly adding to and maintaining the Index.

The present report focuses on the procedures used by SDC to identify and evaluate available student performance measures that can be considered candidate instruments for use in evaluating secondary and postsecondary compensatory education programs. While the procedures used by SDC during the planning phase of the Planned Variations Study were designed to meet specific contractual requirements, the overall approach presented in this report as well as the evaluation results are applicable to other secondary and postsecondary compensatory program evaluations.

Problem

Behavioral scientists who undertake to design and conduct social program evaluations must solve a variety of complex, and at times perplexing, problems if the results of the evaluations are to be informative and useful. Some of the more difficult problems that crop up in the early phases of an evaluation concern the research design and analytic strategies to be employed to answer the research questions. Another area requiring careful attention concerns data collection procedures, including strategies for assuring quality control over the data collected for the evaluation. A third area that presents numerous pitfalls for the evaluation specialist, and the area that serves as the focus of this report, concerns the program objectives and the procedures and instruments used to evaluate how well a program has attained the specified objectives. Tinto and Sherman* note several

*V. Tinto and R. H. Sherman. The Effectiveness of Secondary and Higher Education Intervention Programs. New York: Teachers College, Columbia University, 1974.

of these pitfalls in a recent review of evaluation studies that have focused on secondary and postsecondary educational intervention programs. One problem area has been the rather poor conceptualization and operationalization of program objectives. While this problem may originate with program planners or developers, it has a major impact on the evaluation specialist responsible for assessing how well a program is performing. Without a clear conceptualization of program objectives, the identification of appropriate measurement strategies and valid and reliable instruments that could be used in the evaluation effort is obviously difficult, if not impossible.

The second problem area in previous evaluation studies Tinto and Sherman refer to concerns the somewhat narrow interpretation of formal program outcomes. Typically, previous evaluations have focused on cognitive outcomes (as measured by achievement tests) or on limited student behaviors (such as grade point averages, retention rates, etc.). Important "informal learnings" that have a direct impact on educational and career achievement have been neglected, according to Tinto and Sherman. These "informal learnings" include self-concept, locus of control, and attitudes toward education, among others. Hence, they suggest that the meaning of program outcome should be expanded in any future effort to evaluate secondary and postsecondary programs.

These two problems alone suggest the need for careful consideration, during the planning and design phase of any evaluation, of program objectives, appropriate performance measures, and the technical characteristics of the instruments to be used in the study. That is, it is essential to identify as precisely as possible the data and instrumentation requirements necessary to answer the study's research questions, and to use these requirements to identify potential instruments from which a final selection can be made.

Organization of Report

In the next section, the data and instrumentation requirements, as identified during the study, are presented. Following that section, the criteria used to

evaluate instruments in each student performance domain along with the results and recommendations are presented. The last section discusses some of the problems SDC encountered in performing the evaluation of available performance measures that can be used at the secondary and postsecondary levels. In addition there are three appendixes that provide summary data on achievement tests, study skills tests, and affective tests.

EVALUATION STUDY REQUIREMENTS

In large measure, the structuring of the review and evaluation procedures used to identify and select candidate instruments was influenced by the overall objectives of the evaluation. For the Planned Variations Study these included the following objectives as stated in the Request for Proposals (RFP):

- Determine the longitudinal impact of the planned variations on verbal and mathematics achievement, study skills, and motivation for academic success and further education.
- Determine the differential effectiveness of the models after one, two, or at most three years of implementation.
- Evaluate the cumulative effects of different treatments, under different exposure durations.

Although the RFP mentions other objectives to be addressed in the study, these objectives indicate that initial planning had to incorporate a design that would permit an impact study. That is, the study design, including the outcome measures used, had to be compatible with measuring student performance longitudinally and with deriving student growth measures over different time periods. The student measures selected were to be administered on a pretest/posttest basis, and hence should, if possible, produce stable results from one administration to the next.

In planning for a student-level impact study, the RFP specified four student performance domains that the test review and selection process should be designed around. One domain was denoted "basic skills" and included reading (e.g., vocabulary and comprehension), language arts (e.g., spelling, grammar, syntax), and

mathematics (e.g., arithmetic through algebra). The second student performance domain was study skills. The third domain was broadly denoted "student attitudes," including self-concept, attitudes toward school, and achievement motivation. The last area included in the RFP was "behavioral measures of academic success." Among other "behaviors," the study was to be designed to measure rates of retention in high school, high school graduation, admission to college, and overall grade-point average.* In addition to identifying and evaluating available instruments that measure student performance in these domains, the RFP laid out several criteria to serve as a basis for conducting a trade-off analysis on the available instruments. These additional criteria are:

- The selected instruments should be appropriate to students in grades 10 through 12 and the first two years of college.
- The subtest levels should have some grade level overlap to minimize the possibility of floor and ceiling effects.
- The administration time should be kept at reasonable levels.
- The selected instruments should have minimum ethnic group bias.
- Minority groups should be represented in the standardization sample.
- The tests or instruments should have acceptable test reliability and validity.
- Norms and scales of the instruments should be adequate for the proposed study.
- Test or instrument content should be relevant, interesting, and meaningful to today's minority group and economically disadvantaged students.
- The instruments should be easily administered, scored, and processed.

*The fourth student performance domain, "student behaviors," is not examined in this report. This report focuses only on those student performance measures that have been developed and used in other studies or other program evaluations.

Given these evaluation design and measurement considerations, SDC developed a set of procedures for identifying, screening, and evaluating available tests or instruments that fall within each of the student performance domains. Since somewhat different procedures were used to identify and evaluate instruments in each of the three domains, separate discussions of these procedures are presented in the following sections of the report.

To assist SDC in conducting a student performance measures evaluation, a four member Instrumentation Review Panel (IRP) was organized. The IRP was composed of experts in the fields of achievement testing, attitude measurement, and survey methodology. The role of the IRP was to:

- Review and comment on the procedures SDC used to identify available tests or instruments.
- Review and comment on the test evaluation criteria SDC applied in rating each instrument.
- Suggest other sources for identifying available instruments and propose modifications to the overall instrument evaluation procedures.
- Suggest other student performance areas that should be measured in the evaluation study.

The IRP met in July, 1977, to review the work SDC had completed in identifying and evaluating available instruments. The recommendations that emerged from the meeting culminated in several modifications to the final test review and evaluation work undertaken by SDC in preparing this report. The recommendations SDC concurred with are discussed in the following sections that focus on the three student performance domains. Several recommendations were rejected, either because SDC did not have the resources to implement them or because no way could be found to implement them.

LANGUAGE ARTS, READING, AND MATHEMATICS ACHIEVEMENT TESTS

In the "basic skills" domain, three areas were identified for student performance assessment. They are: (1) mathematics (e.g., arithmetic through algebra), (2) language arts (e.g., spelling, grammar, syntax), and (3) reading (e.g., vocabulary and comprehension). In searching for standardized achievement tests, SDC tried to identify tests that covered these content areas. The Instrumentation Review Panel (IRP) suggested that SDC consider assessing students' actual writing ability instead of their "knowledge" about the mechanics of writing. This recommendation is discussed in more detail at the end of this section.

Screening Criteria and Scoring Procedures

The procedure used to identify and screen achievement tests included two phases. The initial phase identified as many candidate achievement tests as could be found that might fulfill the needs of the study. The sources consulted included the following test evaluations:

Ralph Hoefpner, et al., CSE Secondary School Test Evaluations, Center for the Study of Evaluation, UCLA, 1974.

Ralph Hoefpner, "Achievement Test Selection for Program Evaluation," in Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, edited by Michael Wargo and Donald Green, CTB/McGraw-Hill, 1977.

Measures of Student Growth for the Sustaining Effects Study, System Development Corporation, TM-5693/003/00 (September, 1976)

P. A. Bradley, B. Y. Kooi, and J. M. Petrosko, Selection of Achievement Tests for a Longitudinal Evaluation of the ESAA Pilot Program, Center for the Study of Evaluation, UCLA, 1973.

In addition, major test publishers were contacted to determine whether or not recent revisions in available achievement test batteries had occurred. If such revisions had taken place, the appropriate materials were requested. The achievement test files available at the UCLA Center for The Study of Evaluation were reviewed as well.

In conducting the initial review, achievement tests were screened on three principal criteria, as defined by CSE Secondary School Test Evaluations:

- Examinee Appropriateness: The summary rating by CSE on this criterion included consideration of item difficulty, item appropriateness and relevance, test format, and item coherence.
- Measurement Validity: The summary rating on this criterion included consideration of construct, content, concurrent, and predictive validities.
- Normed Technical Characteristics: The overall rating on this criterion included consideration of test reliability, score graduation, and score distribution.

The above criteria were used to screen out the achievement tests that have inadequate technical characteristics or content limitations that did not meet the needs of the Planned Variations Study. To survive the initial screening, an achievement test battery had to receive an overall rating of "fair" to "good" on each of the three general evaluations criteria.* Altogether, 28 standardized achievement tests were reviewed during the initial phase of the test evaluation process. Six achievement test batteries survived the initial screening, and were evaluated using the criteria discussed below.

The criteria and their associated scoring procedures used in the second phase screening of six achievement test batteries are:

*In evaluating each of the candidate tests during the initial screening, the CSE test evaluations were relied on extensively. In addition to utilizing a quantitative rating scheme in assessing specific features of tests, the CSE evaluators rated each test by using the designations "Poor," "Fair," and "Good" for each overall criterion category.

1. Test Level Appropriateness: The IRP recommended that testing students at "grade level" may be inappropriate. In discussing this issue with personnel from the Bureau of Higher and Continuing Education as well as with Upward Bound and Special Services Project Directors, the general consensus was that project participants are typically at least two years below grade level in reading, writing, and mathematics skills. It was decided, therefore, that high school students would be given a test appropriate for grades 8-10, while college students would receive a test appropriate for grades 11-12. The following rating criterion was defined: If a test battery had a level that is appropriate for either grades 8-10 or grades 11-12, it was rated 1; if a test requires two versions either for grades 8-10 or 11-12, or does not provide adequate grade coverage for either group, it was rated 0.

2. Content Coverage: The test (or tests) to be selected must have, at a minimum, the capability of assessing the students' reading skills (e.g., vocabulary and comprehension), language arts skills (e.g., spelling, grammar, syntax), and mathematics skills (e.g., computation and conceptual reasoning). A test was given a value of 1 for each of the above general content areas covered, and 0 for each one not covered.

3. Completion Time: Inasmuch as students will be completing other types of outcome and background instruments, respondent burden must be minimized. Each subtest was evaluated in terms of how long it takes to complete. The scoring was as follows:

   35 minutes or less = 2

   36-45 minutes = 1

   More than 45 minutes = 0

4. Sample Representativeness: If a publisher had included within the standardization sample adequate representation (in relation to the national population) of minority group students, of students from different geo-

graphical regions, and of students from rural, suburban, and urban communities, then a score of 2 was assigned to the test. If a limited amount of information on the sample was reported and that information indicates the sample was reasonably balanced (at least in terms of the reported characteristics), then a score of 1 was assigned. If inadequate information was given to judge whether or not the sample was balanced, or the information indicated that an unbalanced sample was used, then a score of 0 was assigned.

5. Ceiling and Floor Effects: Ceiling effects were addressed by determining the percentage of students in the standardization sample that received a perfect or near-perfect score on the subtests.* The complete test battery was rated in terms of the subtest that received the highest score. The scores are as follows:

   0% to 5% with perfect or near-perfect score = 2

   6% to 9% with perfect or near-perfect score = 1

   10% or more with perfect or near-perfect score = 0

   Floor effects were assessed by determining the percentage of students that scored at or below the chance level on a subtest. The same percentage ranges as noted above were used to evaluate a subtest's floor effects.

6. Test-Retest Reliability: Each test was assessed in terms of test-retest reliability, with each test's reliability data scored as follows:

   reliability coefficients of .90 and above = 3

   reliability coefficients of .80 to .89 = 2

   reliability coefficients of .79 and below = 1

   no data provided = 0

*A "near-perfect" score was defined as 95%-99% of the test items answered correctly.

7. Test Validity: Each test was assessed in terms of content validity and concurrent validity. If a publisher mounted an elaborate test development process that included identification and classification of measurable objectives, that utilized not only staff specialists but also external consultants (e.g., subject-matter specialists) to develop and review items, that incorporated several pilot tests at various stages of test development, and that reflected item and test revisions throughout the development process, then a score of 2 was assigned to the test. If the development effort was less elaborate and did not make use of external consultants and numerous tryouts or pilot tests, a score of 0 was assigned.

   A scoring procedure for concurrent validity was as follows:

   correlation coefficients of .80 or above = 3

   correlation coefficients of .70 to .79 = 2

   correlation coefficients of .60 to . 69 = 1

   correlation coefficients below .60 or data not provided = 0

8. Debiasing Procedures: If a publisher applied systematic procedures to remove any test items that were biased against minority group students, a score of 1 was assigned. If not, then a score of 0 was assigned.

It should be noted that two additional criteria were dropped from the screening procedures after each of the six achievement tests were initially reviewed. These criteria are: (1) group administration of the test vs. individual student administration, (2) machine scorability of the tests. All of the tests that survived the preliminary screening are group administered and the publishers provide scoring sheets that can be machine scored either optically or magnetically.

Summary of Assessment Results

The six achievement tests assessed in terms of the above criteria are:

Test of Academic Skills, The Psychological Corporation (TASK)

Tests of Academic Progress, Form S, Houghton-Mifflin (TAP/S)

Sequential Tests of Educational Progress, Addison-Wesley Publishing Company (STEP II)

California Achievement Tests, 1970 Edition, CTB/McGraw-Hill (CAT-70)

Comprehensive Tests of Basic Skills, Level 4, Form S, CTB/McGraw-Hill (CTBS/4)

California Achievement Tests, Forms C and D, 1977 Edition, CTB/McGraw-Hill (CAT-77)

Table 1 reports the results from the application of the criteria described in the previous section.* As can be seen from reviewing the total score for each test, the point spread is not very drastic: it ranges from 9 to 13 points. This results, in part, from the fact that an initial screening removed tests that would have scored at the lower total score range. The absence of tests at the upper total score range probably results from using fairly demanding criteria to evaluate tests to be used in this study. The tests reviewed by SDC are, for the most part, very good achievement tests even though no one test meets all the criteria well. To do so, in terms of the needs of the Planned Variations Study, a test would have to score 24 (or close to it) on the rating scale.

The following is a brief summary of the results obtained in applying each criterion:

1. Test Level Appropriateness: All of the test batteries had a test level that would be appropriate for either grades 8-10 or 11-12.

*The achievement test summaries used in this evaluation can be found in the Appendix A.

TABLE 1. Results from an Assessment of Achievement Tests

Assessment Criteria	Scoring Range	TASK	TAP/S	STEP II	CAT-70	CTBS/4	CAT-77
1. Test Level Appropriateness	0-1	1	1	1	1	1	1
2. Content Coverage							
a. Reading	0-1	1	1	1	1	1	1
b. Language Arts	0-1	1	1	1	1	1	1
c. Math	0-1	1	1	1	1	1	1
3. Completion Time							
a. Reading	0-2	1	1	1	0	0	0
b. Language Arts	0-2	1	1	1	0	0	0
c. Math	0-2	1	1	0	0	0	0
4. Sample Representativeness	0-2	2	1	1	1	1	2
5. Evidence of C or F Effects							
a. Ceiling Effects	0-2	2	2	2	2	1	2
b. Floor Effects	0-2	2	0	2	2	0	2
6. Test-Retest Reliability	0-3	0	0	0	0	2	0*
7. Validity Data							
a. Content	0-1	0	1	1	0	1	1
b. Concurrent	0-3	0	2	1	0	0	0*
8. Debiasing	0-1	0	0	0	0	1	1
TOTAL SCORE	0-24	13	13	13	9	10	12

*At the time of this review, SDC did not have access to reliability and validity data on the CAT-77.

2. Content Coverage: All of the tests included subtests that measure the specific skills of interest in this study.

3. Completion Time: Three test batteries (the TASK, TAP/S, and STEP II, except for the mathematics subtest) included subtests that can each be completed within the 35 to 45 minute range. The CAT-70, CAT-77, and CTBS/4 had subtests each requiring more than 45 minutes to complete.

4. Sample Representativeness: Only one test publisher (The Psychological Corporation--Test of Academic Skills) provided sufficient data on the norming sample to make a definitive judgment about sample representativeness. The TASK norming sample was nationally representative in terms of geographical region, minority group, and economic status characteristics. The other test publishers apparently used balanced norming samples at least in terms of one or two student or school characteristics. The CTB/McGraw-Hill norming samples used for the CAT-70 and CAT-77 had some degree of minority group representation, although it was impossible to judge the degree of representation of specific minority groups.

5. Ceiling and Floor Effects: Only one test battery (the CTBS/4) had moderate ceiling effects on some of the subtests. Two test batteries (the TAP/S and the CTBS/4) had fairly strong floor effects on some of the subtests.

6. Test-retest Reliability: Four of the test batteries did not receive this type of reliability examination. The CTBS/4 was studied in terms of its test-retest reliability; the coefficients for subtests fell within the range of .80 to.89. It is impossible to judge, at this time, whether the CAT-77 received such examination.

7. Test Validity:

   a. Content Validity: All of the publishers devoted some attention to ensuring that their respective tests had valid content. However, the development procedures for the TAP/S, the STEP II, and

the CTBS/4 included extensive use of external consultants (e.g., curricular specialists) to evaluate each test's content validity and numerous item revisions were made based on field tryouts.

b. Concurrent Validity: Only two tests (the TAP/S and the STEP II) were exposed to concurrent validity studies. The resulting correlation coefficients fell in the .65 to .79 range.

8. Test Debiasing: Only two tests (CTBS/4 and CAT-77) were exposed to item debiasing procedures used to identify and delete test items that were judged biased against particular minority group students.

Recommendations

Although none of the tests stand out, in terms of the overall ratings, as a superior candidate over the others, there were important differences on specific criteria. For example, three of the six tests received the same overall rating which would suggest any one of them would be an excellent candidate for use in secondary or postsecondary evaluation studies. However, each differed on several important evaluation criteria. While the norming sample for the TASK was rated superior to the norming samples for the TAP/S and STEP II, the validity data for the latter two indicate they had excellent validity characteristics as compared to the TASK. Also compounding the problem of deciding which battery might be the best choice was the fact that CTB/McGraw-Hill apparently was the only test publisher (of the four whose tests are reviewed here) that provided published information on the application of statistical debiasing procedures. While all of the major test publishers indicated they applied such procedures to their tests, CTB/McGraw-Hill was the only one that reported such information.

One test that might stand out as the superior candidate was the CAT-77. Unfortunately, complete technical data on its validity and reliability characteristics were not available at the time this report was prepared. The CAT-77 was an extensively revised version of the CAT-70 and included many attractive features that recommend it for use in evaluating Upward Bound or Special Services-type compensatory education programs. Some of the more salient features are:

1. The CAT-77 has been extensively scrutinized for items that might be biased against particular ethnic or racial groups. Content and statistical analyses were performed, culminating in the removal of items that seemed to be biased against minority students.
2. The norming sample characteristics are good, and the CAT-77 includes fall and spring norm data.
3. The CAT-77 results can be reported in both norm-referenced and criterion-referenced formats.
4. "Locator tests" are available that can be used to determine, prior to administration of the entire battery, the functional skill levels of the students and which test level would be appropriate. "The locator tests consist of short vocabulary and mathematics sections that have been shown to be reliable in matching students who are in the same grade with different levels of CAT C and D."*

Given these features, the CAT-77 would probably be the best achievement test battery for use in a secondary or postsecondary program evaluation. However, until reliability and validity data are available from CTB/McGraw-Hill, a definitive recommendation concerning the CAT-77 cannot be made.

Of the five remaining achievement test batteries, the STEP II would probably be the best test to use in evaluting secondary and postsecondary compensatory education programs. SDC based this recommendation on several considerations. In general, the STEP II has relatively strong technical characteristics as defined by the test rating procedures discussed early. There is no evidence of ceiling or floor effects in the normative data, and the battery has fairly

*Quoted from Test Coordinator's Handbook for CAT C and D (draft).

strong content and concurrent validity data. The STEP II has reasonable completion time requirements and acceptable norming sample characteristics. In addition, the battery includes a test level that can be used for grades 13 and 14 if required. The only other battery that goes beyond high school is the TASK, which is recommended for use among junior college freshmen only.

The final consideration SDC made in deciding to recommend the use of the STEP II stems from comments made by the Instrumentation Review Panel. After reviewing five of the six achievement tests (CAT-77 was not available at the time of the meeting), the IRP suggested that the STEP II battery seemed to be the best test available. The panel members indicated that this battery seemed to be superior in content, layout, the skill areas covered, as well as offering very strong features in the area of validity and psychometric characteristics.

Assessing Writing Skills

One of the student performance areas to be assessed in PVS was writing skills. All of the achievement test batteries reviewed in this report included subtests that measure knowledge of proper writing techniques, principles, and mechanics (that is, an indirect measure of writing skills). However, the IRP suggested that SDC consider measuring directly a student's actual writing ability by having him write a short essay. The student essays would then be scored, with the results used to assess students' overall writing abilities.

SDC examined the IRP suggestion carefully in terms of its feasibility as well as in terms of the additional information to be obtained from administering an essay examination. The Educational Testing Service has done extensive research on the advantages and disadvantages associated with using essay versus multiple-choice examinations. One study was conducted by H. M. Breland and

J. L. Braucher,* the results of which are relevant to the IRP suggestion. The Breland and Braucher study involved four institutions with large entering freshmen classes. The data collected included: (1) Test of Standard Written English (TSWE), an objective test in multiple-choice format. (2) the SAT completed prior to entering college; and (3) 20-minute written essays completed when the students entered college, in the spring of their freshman year, and the following fall. One analysis focused on the relationships between performance on the TSWE and performance on the essays. The results, as reported by Breland and Braucher, were:

a. The correlation between the TSWE (indirect measure of writing skills) and the essays (direct measure) was .76 and was statistically significant.

b. The TSWE appeared to predict later writing performance better than a 20-minute essay.

These results suggest that the additional effort and evaluation costs associated with using an essay test to measure writing ability would not be repaid in better evaluation data on students' writing performance. SDC concluded, therefore, that a multiple-choice writing ability test would be adequate for secondary or postsecondary evaluation studies.

---

*H. M. Breland and J. L. Braucher, "Measuring Writing Ability," a paper presented at the Annual Meeting of the American Educational Research Association, New York, April 1977. Also see H. M. Breland, A Study of College English Placement and the Test of Standard Written English. Princeton, New Jersey: Educational Testing Service, 1977.

STUDY SKILLS TESTS

Assessment of program impact on students' study skills is another outcome area examined by SDC during the Planned Variations Study. The approach taken by SDC to this student performance domain included consideration of two primary components. One component of study skills focused on the student's knowledge of proper study techniques and practices. The second component focused on the student's actual study habits; that is, how the student actually goes about studying. While knowledge of proper techniques is important, SDC feels that how the student actually studies is at least equally important. The search for appropriate standardized study skills tests was guided by this consideration.

Screening Criteria and Scoring Procedures

In conducting the initial review, the following criteria were used to screen out tests that did not warrant further consideration:

- Examinee Appropriateness: The summary rating on this criterion included consideration of item difficulty, item appropriateness and relevance, test format, and item coherence.
- Measurement Validity: The summary rating on this criterion included consideration of construct, content, concurrent, and predictive validities.
- Normed Technical Characteristics: The overall rating on this criterion included consideration of test reliability, score graduation, and score distribution.

Initially, a study skills test had to receive an overall rating of "fair" to "good" by CSE evaluators on each of the three general evaluation criteria. However, this screening requirement had to be relaxed in order to retain tests for the second phase screening. The relaxed requirement was that a test had to receive a "fair" to "good" rating on two of the three criteria. In addition, a test had to include items that measured fairly comprehensively the skills and knowledge required to study effectively. Many standardized study skills tests are designed to measure a limited number of skills, such as how to use a library

or a reference book. For compensatory intervention evaluations, it is necessary to utilize a study skills test that measures a variety of skills, knowledge, and attitudes that pertain to effective study behavior.

During the initial screening, 14 tests were identified. Three tests survived the initial screening. The criteria and their associated scoring procedures used for the second screening are:

1. Content Coverage: Ideally, the study skills test (or tests) selected should have two general components. One component includes items to measure what the student knows regarding correct outlining procedures, map-reading procedures, and so on. This is the cognitive skill component. The other test component includes items to measure what the student actually does when studying. This is the study habits component. If a test has both components, a score of 2 was assigned. If it has only one component, it was scored 1. If neither component is measured by the test, a score of 0 was assigned.

2. Test Level Appropriateness: If a test is appropriate for either grades 8-10 or 11-12, it was given a rating of 1. If it was inappropriate for either level, it was scored 0 on this criterion.

3. Sample Representativeness: If a publisher has included within the standardization sample adequate representation (in relation to the national population) of minority group students, or students from different geographical regions, and of students from rural, suburban, or urban communities, then a score of 2 was assigned to the test. If a limited amount of information on the sample is reported and that information indicates the sample is reasonably balanced (at least in terms of the reported characteristics), then a score of 1 was assigned. If inadequate information was given to judge whether or not the sample is balanced, or the information indicates that an unbalanced sample was used, then a score of 0 was assigned.

4. Ceiling and Floor Effects: Ceiling effects were addressed by determining the percentage of students in the standardization sample that received a perfect or near-perfect score on the subtests.* The complete test battery was scored in terms of the subtest that received the highest score. The scores are as follows:

   0% to 5% of the students received a perfect or near-perfect score = 2

   6% to 9% with perfect or near-perfect score = 1

   10% or more with perfect or near-perfect score = 0

   Floor effects were assessed by determining the percentage of students that scored at or below the chance level on a subtest. The same percentage ranges as noted above were used to evaluate a subtest's floor effects.

5. Test Reliability: Each test was assessed in terms of test-retest reliability, internal consistency reliability, alternate form reliability, and inter-level reliability. Each type of reliability data was scored as follows:

   reliability coefficients of .80 and above = 3

   reliability coefficients of .70 to .79 = 2

   reliability coefficients of .60 to .69 = 1

   reliability coefficients below .60 or no data provided = 0

6. Test Validity: Each test was assessed in terms of content validity, concurrent validity, and predictive validity. If a publisher mounted an elaborate test development process that included identification and classification of measurable objectives, that utilized not only staff specialists but also external consultants (e.g., subject-matter specialists) to develop and review items, that incorporated several field tryouts or pilot tests at various stages of test development, and that

*A "near-perfect" score was defined as 95%-99% of the test items answered correctly.

reflected item and test revisions throughout the development process, then the test was scored 1 on content validity. If the development effect was less elaborate and did not make use of external consultants and numerous tryouts or pilot tests, a score of 0 was assigned.

The scoring procedure for concurrent validity is as follows:

correlation coefficients of .80 or above = 3
correlation coefficients of .70 to .79 = 2
correlation coefficients of .60 to .69 = 1
correlation coefficients below .60 or data not provided = 0

The same coefficient ranges were used to assess a test's predictive validity.

7. Test Administration Characteristics: Several aspects of each test were evaluated within this criterion category. If a test can be administered in a group setting (instead of individual administration), it was scored 1. Otherwise, it was scored 0. If a test has an easy-to-use answer sheet, it was scored 1. If a test has a fairly complicated answer sheet format, it was scored 0. Otherwise, a score of 1 was assigned. Last, the following scoring procedure was used to evaluate each test in terms of test administration time:

Less than 25 minutes to complete = 2
25-35 minutes to complete = 1
More than 35 minutes to complete = 0

If insufficient information is provided to rate a test on any of the above administrative criteria, a 0 was assigned to the test.

Summary of Assessment Results

The three study skills tests assessed using the above criteria are:

Effective Study Test, Effective Study Materials, Inc. (EST)

Survey of Study Habits and Attitudes, The Psychological Corporation. (SSHA)

The Study Skills Counseling Evaluation, Western Psychological Services. (SSCE)

Although the EST and SSHA are marketed by separate organizations, they were developed, at least during the early stages, jointly; further, the principal author of the EST also contributed to the development of the SSHA. One apparent objective of this joint development effort was to develop tests that would measure both knowledge about proper study skills and actual study habits.

Table 2 reports the test assessment results.* The following is a brief summary of the rating results.

1. Content Coverage: None of the tests reviewed here included items that measure both the cognitive and study habits components of study skills. The EST includes items that measure the student's knowledge of correct study techniques, while the SSHA and SSCE measure the student's study habits.

2. Test Level Appropriateness: All three tests were developed for grades 10-14. The SSCE does not have normative data by grade (only by institutional levels, e.g., high school, college, etc.). The EST and SSHA do not have norms for college sophomores.

3. Sample Representativeness: None of the normative samples seems balanced in terms of ethnic or regional group representation.

4. Ceiling and Floor Effects: The SSHA appears to be the only test with floor effects. None of the tests has ceiling effects.

*The test summaries can be found in Appendix B.

TABLE 2. Results From an Assessment of Study Skills Test

Assessment Criteria	Scoring Range	EST	SSHA	SSCE
1. Content Coverage				
a. Cognitive Skills	0-1	1	0	0
b. Study Habits	0-1	0	1	1
2. Test Level Appropriateness	0-1	1	1	1
3. Sample Representativeness	0-2	0	0	0
4. Ceiling or Floor Effects				
a. Ceiling Effects	0-2	2	2	2
b. Floor Effects	0-2	2	0	2
5. Test-Retest Reliability	0-3	0	3	3
6. Validity Data				
a. Content	0-1	1	1	0
b. Concurrent	0-3	0	0**	0
c. Predictive	0-3	1	0	0
7. Administration Characteristics				
a. Administration Context	0-1	1	1	1
b. Answer Sheet Usability	0-1	1*	1	1
c. Test Administrator Training Requirements	0-1	1	1	1
d. Administration Time	0-2	0	1	2
TOTAL SCORE	0-24	11	12	14

*The answer sheet provided by the publisher can be hand-scored only. A scoring key is provided and the publisher estimates that it takes about 90 seconds to score each test.

**This zero rating is based on the correlation coefficients reported in the EST Manual. See Table 3 in the EST evaluation summary for the specific correlation coefficients relating EST scores with SSHA scores.

5. Test-retest Reliability: While the EST has not been studied for its test-retest reliability, both the SSHA and SSCE have very stable test-retest reliability characteristics.

6. Test Validity:

   a. Content Validity: The development procedures used for the EST and SSHA included guidelines to ensure the content validity of test items. No such procedures were reported in the SSCE manual.

   b. Concurrent Validity: Although all of the tests reported data relevant to this criterion, the correlation coefficients were consistently below .60.

   c. Predictive Validity: The correlation coefficients relating GPA to EST scores at the college level are below .60. At the high school level, the correlation coefficients cluster (on the average) around .60 (see Table 2 in the EST summary). The correlation coefficients relating SSHA scores to the students' GPA are all below .60. Similarly, the correlation coefficients for the SSCE are below .60.

7. Administration Characteristics:

   a. Administration Context: All tests can be administered in a group setting.

   b. Answer Sheet Usability: The three tests have an easy-to-use answer sheet. The EST answer sheet can only be hand-scored.

   c. Test Administrator Training Requirements: Neither test requires specially-trained administrators.

   d. Test Administration Time: The EST requires from 40 to 45 minutes to complete, while the SSHA requires between 25 and 35 minutes to complete. The SSCE can be completed in about 20 minutes.

Recommendation

Similar to the achievement tests reviewed earlier, none of the study skills tests stands out from the others. However, the SSCE did receive a total rating above the other two tests, and does have several features that make it an attractive candidate for use in evaluations. It is designed for use at both the high school and college levels, and the normative data indicate that very few of the students received a perfect score or scored below the chance level. Another attractive feature is that the test can be completed within 20 minutes, while the other two require somewhat more time to complete.

Unfortunately, none of the tests reviewed here measure both cognitive skills and study habits. If it were feasible, in terms of respondent burden and cost considerations, to administer the SSCE (a measure of study habits) and the EST (a measure of cognitive skills), SDC would recommend such a strategy. However, if this is not a feasible option, then the SSCE seems to be the best study habits measure available for use at the secondary and postsecondary levels.

STUDENT AFFECTIVE MEASURES

The third student performance domain examined by SDC was designated "student attitudes," and originally included three specific areas of attitude measurement: (1) the student's educational self-concept, (2) the student's orientation to school, and (3) the student's educational achievement motivation. In assessing changes in a student's affective behavior, the primary emphasis was on measuring attitudes pertaining specifically to school and the learning process. Hence, in identifying candidate tests, SDC searched for those that measured, for example, a student's self-concept in the educational or school setting, and not self-concept in general. This guideline applied as well to the other affective outcome areas.

It should also be noted that the IRP recommended that SDC include a locus of control measure with the other three affective measures. SDC was able to identify and evaluate several candidate measures of locus of control. The results of this effort are reported in this section along with the results from the other affective areas.

Screening Criteria and Scoring Procedures

Similar to the procedure used to review achievement tests, affective tests were reviewed by following a two-step procedure. The initial step involved the identification of as many candidate tests for each affective outcome area as could be found. The sources consulted included the following evaluations of many available affective tests:

Ralph Hoepfner, et al., CSE Secondary School Test Evaluations (1974).

John Robinson and Phillip Shaver, Measures of Social Psychological Attitudes (Revised Edition, 1973).

K.T. Chun, S. Cobb, and J.R.P. French, Jr., Measures for Psychological Assessment (1975).

Ruth Wylie, The Self-Concept (Revised Edition, 1974).

In addition, SDC staff reviewed the affective test materials available at the Center for the Study of Evaluation in order to identify any new or old tests that had not been uncovered by reviewing the above sources. Consultants to PVS also proved to be an invaluable resource, especially in identifying affective measures developed and used by university researchers.

For initial screening, each test was assessed in terms of three primary criteria:

1. Grade-Level Appropriateness: Tests were dropped from further consideration if they were not designed for or used at least at the high school level. (Early in the review process it became apparent that very little work has been done on measuring the affective characteristics of college students. Hence, the initial screening identified appropriate high school tests, and these were assessed in the final screening step in terms of being appropriate at the college level.)

2. Content Appropriateness: To receive further consideration, a test had to be a measure of one of the affective areas of interest in this study, and include items that seemed to measure the individual's affective characteristics in the school context.

3. Administrative Usability: A test was dropped from further consideration if it posed interpretation problems or scoring difficulties, or could not be administered to groups of students, or required too much time to complete (50 minutes or more).

The initial screening identified 45 tests that were classified as measures of self-concept, 25 that were classified as measures of achievement motivation, 30 that were classified as measures of school orientation, and 12 that were classified as measures of locus of control. From the initial list of instruments, three self-concept instruments, six school orientation instruments, and three locus of control instruments were identified for additional in-depth evaluation during the second stage of the review process. It should be stressed that requiring candidate instruments to measure various dimensions of a student's

attitudes to school and the learning process places a severe constraint on locating a large pool of instruments. This criterion alone accounts for the substantial reduction in the number of instruments suitable for evaluation in the second phase. A second factor that contributed to the reduction in the number of suitable instruments is the very poor psychometric characteristics of many available affective instruments.

None of the achievement motivation instruments initially reviewed was found suitable for additional in-depth evaluation. Similar to many of the other affective tests, the achievement motivation tests did not have appropriate content or had severe limitations in terms of not being amenable to group administration. Most of the instruments initially reviewed also required fairly elaborate scoring procedures that make their use inefficient in large-scale evaluations. However, SDC subsequently uncovered one achievement motivation instrument that seemed appropriate for use in large-scale secondary or postsecondary evaluations. The characteristics of this instrument are discussed below.

The criteria and their associated scoring procedures used during the second stage are:

1. Test Content: As noted above, a key consideration in reviewing affective instruments was whether a test was designed to measure the student's affective characteristics (e.g., self-concept, school orientation, locus of control, achievement motivation) in general, or in the school context and student role. The following scoring procedure was applied in evaluating tests for each of the areas of attitude measurement under consideration in PVS: If most (more than 60%) of the items address one of the specific affective areas of interest, the test was scored 2. If approximately 20% to 60% of the items address one of the specific areas, the test was scored 1. If less than 20% of the items are devoted to one of the specific areas, the test was scored 0.

2. Test Level Appropriateness: If it appeared, either on the basis of the developer's recommendation or on the basis of reading the items, that a test could be used either for grades 8-10 or for grades 11-12, it was scored 1. If it was inappropriate for either level, it was scored 0.

3. Sample Characteristics: If a test developer or researcher included within the test administration sample students from different socio-economic strata (e.g., low, middle, high income groups) or from different ethnic or racial groups, the test was scored 1 on this criterion. If the sample appeared to be unbalanced or very specialized, or if insufficient data were provided to make a judgment, the test was scored 0.

4. Ceiling and Floor Effects: In attempting to assess the tests in terms of floor or ceiling effects, two strategies were used. One strategy was to follow the same procedures used to assess achievement tests on this criterion. However, developers of affective tests normally do not undertake standardization of their instruments, hence percentile distributions usually are not available. A second strategy was to search for data that had been collected using each instrument. If the reported data included a sample mean and standard deviation, then these statistics were used to assess the presence or absence of ceiling or floor effects. If the data suggested that such effects are not present, the test was scored 1. If the data suggested that such effects are present, a score of 0 was assigned. If the data were insufficient to make a judgment, the test was scored 0 on this criterion also.

5. Test Reliability: Each test was reviewed in terms of test-retest reliability, internal consistency reliability, alternate form reliability, and inter-level reliability. Each type of reliability data was scored as follows:

   reliability coefficients of .80 and above = 3
   reliability coefficients of .65 to .79 = 2
   reliability coefficients of .50 to .64 = 1
   reliability coefficients below .50 or no data provided = 0

   Whenever a test has reliability data on several subdimensions, and an overall reliability coefficient is not reported, an average reliability coefficient was calculated. In making this calculation, the Spearman-Brown formula was applied using Fisher *z* values for the reliability coefficients.

6. Test Validity: Each test was evaluated in terms of content and concurrent validity. If a test developer indicated that this was a development criterion and that various revisions occurred as a result of pilot tests and test reviews, a score of 1 was assigned to the test. If the development process did not include content validity as a development criterion or if the technical information reviewed was insufficient to make a judgment, the test was scored 0 on this criterion.

   To assess the concurrent validity of a test, it is necessary to have correlation data reflecting the extent to which the test results are associated with criterion measures obtained at the same time as the test results. The scoring procedures used to assess concurrent validity of a test is as follows:

   correlation coefficients of .80 or above = 3
   correlation coefficients of .65 to .79 = 2
   correlation coefficients of .50 to .64 = 1
   correlation coefficients below .50 or no data provided = 0

7. Test Administration Characteristics: Several aspects of each test were evaluated within this criterion category. If a test can be administered in a group setting (instead of individual administration), it was scored 1. Otherwise, it was scored 0. If a test has an easy-to-use answer sheet, it was scored 1. If a test has a fairly complicated answer sheet format, it was scored 0. If a test requires specially trained test administrators, it was scored 0. Otherwise, a score of 1 was assigned. Last, the following scoring procedure was used to evaluate each test in terms of test administration time:

   less than 15 minutes to complete = 2
   15 to 25 minutes to complete = 1
   more than 25 minutes to complete = 0

   If insufficient information is provided to rate a test on any of the above administrative criteria, a 0 is assigned to the test.

Summary of Results and Recommendations

Assessment data on the instruments that supposedly measure educational self-concept, school orientation, locus of control and achievement motivation are presented in Table 3.

Self-Concept Measures

Among the three educational self-concept measures reviewed in depth, the "Self-Concept of Ability Scale" (SCAS) received the highest overall rating. It would be an especially good instrument to use in secondary and postsecondary program evaluations in part because of its educationally-relevant content and its test-retest stability. The SCAS is also attractive because of its length; it can be completed in less than five minutes. SDC recommends its use in measuring secondary and postsecondary students' self-concept.

School Orientation Measures

The six instruments reviewed to measure school orientation did not differ much on the rating criteria. The only obvious differences that did occur were in terms of the test content, test-retest reliability, and administration time criteria. As the data in Table 3 indicates, the IOX "Subject Area Preference" (SAP) scale had the highest rating on the test-retest reliability criterion. On the other hand, the IOX "School Sentiment Index" (SSI) and Bills' "Feelings About School" had the most appropriate content (to measure school orientation) and require the least amount of time to administer.

Although the SAP recieved the highest total score on all criteria, SDC recommends the FAS for use in secondary or postsecondary program evaluations attempting to measure students' orientations or attitudes toward school. This recommendation is made for several reasons. First, it did receive a higher overall rating than the SSI. This difference resulted from the FAS's having excellent score distributions. Second, all of the items in the FAS pertain specifically to the school and to the people and activities making up the school environment. Moreover, the factor analyses performed on FAS data produced two factors that

TABLE 3. Results From an Assessment of Affective Tests (1 of 3)

Assessment Criteria	Scoring Range	Self-Concept Measures		
		SAI (IOX)	For All I Know (IOX)	SCAS
1. Test Content	0-2	1	2	2
2. Level Appropriateness	0-1	1*	1*	1
3. Sample Characteristics	0-1	1	1	1
4. Ceiling or Floor Effects				
a. Ceiling Effects	0-1	0	0	0
b. Floor Effects	0-1	0	0	0
5. Test-Retest Reliability	0-3	1	0	2
6. Validity Data				
a. Content	0-1	1	1	1
b. Concurrent	0-3	0	0	0
7. Administration Characteristics				
a. Administration Context	0-1	1	1	1
b. Answer Sheet Usability	0-1	1	1	1
c. Test Administrator Training Requirements	0-1	1	1	1
d. Administration Time	0-2	2	2	2
TOTAL SCORE	0-18	10	10	12

*This test was developed for use in high schools and the available data on the test were gathered at the high school level. However, with minor word changes (e.g., substitute "instructor" for "teacher," "students" for "kids," etc.) this test would be appropriate for use at the college level.

TABLE 3. Results From an Assessment of Affective Tests (2 of 3)

Assessment Criteria	School Orientation Measures					
	SSI (IOX)	Imagine That (IOX)	High Schoo on TV (IOX)	What Would Happen (IOX)	SAP (IOX)	FAS
1. Test Content	2	1	0	1	1	2
2. Level Appropriateness	1*	1*	1	1*	1*	1**
3. Sample Characteristics	1	1	1	1	1	1
4. Ceiling or Floor Effects						
a. Ceiling Effects	0	0	0	0	0	1
b. Floor Effects	0	0	0	0	0	1
5. Test-Retest Reliability	0	1	1	1	3	0
6. Validity Data						
a. Content	1	1	1	1	1	1
b. Concurrent	0	0	0	0	0	0
7. Administration Characteristics						
a. Administration Context	1	1	1	1	1	1
b. Answer Sheet Usability	1	1	1	1	1	1
c. Test Administrator Training Requirements	1	1	1	1	1	1
d. Administration Time	1	2	2	2	2	1
TOTAL SCORE	9	10	9	10	12	11

*This test was developed for use in high schools and the available data on the test were gathered at the high school level. However, with minor word changes (e.g., substitute "instructor" for "teacher," "students" for "kids," etc.) this test would be appropriate for use at the college level.

**This test would require a moderate amount of item revision to be used at the postsecondary level.

Table 3. Results From An Assessment of Affective Tests (3 of 3)

Assessment Criteria	Scoring Range	Locus of Control			Achievement Motivation
		Multi-Dimensional I.E.	Rotter's I.E.	NLS I.E.	Motivational Style Scale
1. Test Content	0-2	1	1	2	1
2. Level Appropriateness	0-1	1	1	1	1
3. Sample Characteristics	0-1	0	1	1	0
4. Ceiling or Floor Effects					
a. Ceiling Effects	0-1	0	0	0	1
b. Floor Effects	0-1	0	0	0	1
5. Test-Retest Reliability	0-3	0	2	0	0
6. Validity Data					
a. Content	0-1	0	0	0	0
b. Concurrent	0-3	0	0	0	0
7. Administration Characteristics					
a. Administration Context	0-1	1	1	1	1
b. Answer Sheet Usability	0-1	1	1	1	1
c. Test Administration Training Requirements	0-1	1	1	1	1
d. Administration Time	0-2	1	1	1	1
TOTAL SCORE	0-18	6	9	8	8

can easily be interpreted as reflecting negative and positive attitudes toward school.* Third, the FAS has been completed by more than 20,000 students, and the normative data from this sample are available for comparison with FAS data from other secondary samples.

Locus of Control Measures

Although Table 3 presents evaluative data on three locus of control measures that could be used in a secondary or postsecondary program evaluation, none of the instruments or scales is very satisfactory. The locus of control items included in the National Longitudinal Study came the closest to measuring students' locus of control in the educational context. However, the NLS scale's psychometric characteristics are generally poor and the scale is composed of only five items. An instrument having better psychometric characteristics and including more items would probably be more appropriate for a secondary or postsecondary evaluation.

The other instruments not only had fairly weak psychometric properties but also did not include sufficient items to tap locus of control in the educational setting. SDC concludes, therefore, that none of these instruments is adequate for use in a secondary or postsecondary evaluation designed to measure educational locus of control. Moreover, any future large-scale evaluation, designed to tap educational locus of control, sponsored by USOE should give serious consideration to allocating project resources to the development and validation of a secondary or postsecondary locus of control measure.

---

*Appreciation is expressed to Dr. Robert E. Bills (College of Education, The University of Alabama) for providing additional data on the FAS instrument.

Achievement Motivation Measures

As mentioned earlier, only one of the identified achievement motivation instruments seems appropriate for use in large-scale secondary or postsecondary educational evaluations. Although the "Motivational Style Scale" (MSS) has not been used extensively, the limited data available on its application suggest that it would be suitable for such evaluations.* The MSS can be administered in a group setting, requires 15 to 20 minutes to complete, and can be efficiently scored. Although the instrument was designed for use at the college level, it could be modified for use at the secondary level. Unfortunately, the MSS has not been used with a sample of minority students, hence there is no way to determine how minority students would score on the subscales. Similarly, there are no data available on its test-retest reliability or its concurrent validity. Nevertheless, SDC recommends its use in measuring achievement motivation among students at the secondary or postsecondary level.

PROBLEMS ENCOUNTERED IN THE INSTRUMENTATION PROCESS

The purpose of this section is to present several problems that were encountered by SDC in conducting a comprehensive search for cognitive and affective measures that might be used in evaluating secondary or postsecondary compensatory education programs. In presenting some of these problems, the hope is that researchers or program evaluators undertaking similar evaluations will be aware of these problems and will be able to make appropriate accommodations to them.

Content of Achievement Tests

A major problem encountered by SDC concerned the identification of achievement tests that measured the curricular content and objectives of extant Upward Bound and Special Services projects that were to serve as comparisons to the programs SDC developed. Although extant Upward Bound and Special Services projects are required to provide services in the areas of basic skills development, it was very difficult for SDC to judge whether or not any of the available standardized tests included content consistent with what was taught in those projects. For programs developed, implemented, and evaluated by the same agency or organization, this judgment seemingly would be less difficult to make. However, an external evaluator responsible for evaluating a program that has been developed by another organization is always placed in the difficult position of questioning the content appropriateness of the achievement test selected for the evaluation. The Instrumentation Review Panel suggested that this has been a major problem with previous secondary and postsecondary education evaluations and that it might account, in part, for the fact that no program effects were found. Short of conducting curricula surveys at the secondary and postsecondary level, the results from which would be used to develop a new standardized achievement test, SDC concluded that using available standardized achievement tests was the best solution to this problem. Evaluations carried out in the future would be strengthened by including a provision for determining the extent of overlap between the content of standardized achievement tests and the content of intervention-program instructional services.

Identifying Affective Measures

A second problem encountered by SDC concerned the identification of affective measures that met the requirements of a systematic evaluation. Despite the rather substantial number of self-concept, achievement motivation, school orientation, and locus of control measures that are readily available, very few had a majority of the features that were required. In general, only a small number of the affective measures identified included educationally-relevant content. This was especially problematic with achievement motivation and locus of control measures. An additional problem concerned the rather limited development of normative, reliability, and validity data on most of the affective measures. Among those measures that did have some validity or reliability data, the data were typically collected on very specialized samples of respondents. Given this situation with affective measures, it would probably be in the interest of future large-scale secondary or postsecondary educational program evaluations to devote sufficient resources to the development and validation of the affective measures to be used in the evaluation.

Locating New Affective Instruments

Somewhat related to the second problem was the problem SDC encountered in locating affective measures developed by or used by university researchers. Unlike standardized achievement batteries, most of the development work in the affective domain has been done by university researchers. And while some effort has gone into compiling lists of available affective measures, most of these lists are very selective and/or dated. Any future effort to identify available affective measures should include a procedure for identifying current developmental work that is in progress and an attempt made to benefit from that current work.

Examining Tests for Bias

Considerable concern has been expressed in recent years over the potential bias in evaluation instruments, bias towards ethnic minorities. As was pointed out in the reviews of various instruments, SDC found that few instruments were subjected to examination for ethnic-group biases by their developers. Given the possibility of such bias, two alternatives are available. First, instruments chosen for use in an evaluation can be limited to those that were debiased during their development. (This would restrict severely the number of candidate tests.) Second, an instrument selected for inclusion in an evaluation can be examined by members of ethnic-minority groups to identify biased items; the latter would then be excluded from any analyses. (This method has been used in some evaluation studies, and requires considerable time and effort on the part of the evaluator.) Neither solution is altogether satisfactory, but either could be implemented in future evaluations of secondary and postsecondary compensatory programs.

APPENDIX A

ACHIEVEMENT TEST SUMMARIES

Name of Test: Test of Academic Skills (Stanford TASK)

Publisher: The Psychological Corporation

Test Levels: Level I: Grades 8-10 (Forms A and B)
Level II: Grades 11-12 and junior/community college freshmen (Forms A and B)

1. Test Content, Number of Items, and Completion Time by Test Level

a. Reading (78 items; 40 minutes)

	No. of Items Level I	No. of Items Level II
Global Meaning	6	5
Meaning of Explicit Detail	18	10
Meaning of Implicit Detail	10	12
Meaning from Context	1	3
Inferences and Logical Analysis	6	12
Meaning from Context in a Modified Cloze Technique Situation	9	9
Word Meaning	27	27

b. English (69 items; 40 minutes)

Learning Skills	15	15
Usage Conventions	21	21
Spelling	15	15
Sentence Sensitivity	6	6
Paragraph Arrangement	12	12

c. Mathematics (48 items; 40 minutes)

Numbers, Symbols and Sets	7	5
Number, Properties and Operations		
• Whole Numbers	8	7
• Common and Decimal Fractions	11	10
• Integers and Exponents	3	2

	No. of Items Level I	No. of Items Level II
Mathematical Sentences	3	4
Geometry and Measurement	2	5
Ratio and Percent	5	4
Graphs, Probability, and Statistics	5	8
Mathematical Reasoning	4	3

2. Norm Data Available For: Level I, Grades 8-10

   Level II, Grades 11-freshman year of junior/ community college

Fall (1972) norm data are presented for both test levels.

3. Representativeness of Standardization Sample

Comparison of TASK Standardization Sample With 1970 Census Data on Selected Demographic Characteristics (Source: Test Publisher's Manual.)

Characteristic	TASK	1970 Census
1. Geographic Region:		
a. Northeast	20%	24%*
b. North Central	34%	28%*
c. Southeast	30%	22%*
d. West	16%	26%*
2. Median Family Income	$9,294	$9,590
3. Median Years of Schooling of Adults in the Community	12.1	12.1
4. Percent of major ethnic minorities in Public, Secondary Schools:		
a. Negro (Black)	16%	11%*
b. Hispanic (Mexican, Puerto Rican, Cuban)	5%	5%*

*Total population

4. Evidence of Ceiling and Floor Effects

   a. One to two percent of the students at each grade level received a perfect or nearly perfect score on each of the subtests.

   b. Typically, only 1 to 4 percent of the students at each grade level scored below the chance level on each subtest.

5. Reliability Data

   a. Test-Retest Reliability: Not assessed by publisher.

   b. Internal Consistency Reliability: Internal consistency of the tests was measured by split-half estimates based on odd-even items (corrected by Spearman-Brown formula) and by the Kuder-Richardson Formula 20. Reliability coefficients range between .92 and .96 for Reading, English, and Mathematics.

   c. Alternate Form Reliability: No data are provided by the publisher in the TASK technical manual.

   d. Inter-Level Reliability: No data are provided by the publisher in the TASK technical manual.

6. Validity Data

   a. Content Validity: Content specialists at the Psychological Corporation identified instructional objectives (by reviewing curricular guides) to be measured by each test. Tests were created and classified into test objectives categories. No evidence that external consultants were used. Some pilot testing occurred throughout test development.

   b. Concurrent Validity: No data are provided by the publisher in the TASK Technical Manual.

7. Item Debiasing Procedures Followed

   There is no indication in the TASK technical manual that item debiasing procedures were used.

Name of Test: Tests of Academic Progress, Form S (TAP/S)

Publisher: Houghton-Mifflin

Test Level: Form S: Grades 9-12

1. Test Content, Number of Items, and Completion Time by Test and Grade Level

	Number of Items by Grade			
	Gr. 9	Gr. 10	Gr. 11	Gr. 12
a. Reading (45 minutes)	61	64	64	67
- Comprehension				
- Inference				
b. Composition (45 minutes)	64	65	65	66
- Capitalization				
- Punctuation				
- Grammar and Usage				
- Organization				
- Spelling				
c. Literature (45 minutes)	55	56	58	60
- Context understanding				
- Understanding content				
- Techniques				
- Inference				
d. Mathematics (45 minutes)	48	48	48	48
- Math principles				
- Math concepts				
- Applications				

2. Norm Data

Available for Form S, Grades 9-12. Norm data for the TAP/S were collected in October and November, 1970; the TAP/S was standardized in coordination with the standardization of the Iowa Tests of Basic Skills and the Cognitive Abilities Test.

3. Representativeness of Standardization Sample

Population and Weighted Sample Distribution
by Community Size Category
(Source: Test Publisher's Manual.)

Community Size Category	Population Percent	Sample Weighted Percent
1,000,000+	9.2	9.48
250,000+	11.5	11.36
100,000+	6.7	8.47
50,000+	8.3	8.61
25,000+	8.5	8.88
10,000+	10.2	11.55
Counties	45.6	41.65
TOTALS	100%	100%

Population and Weighted Student Sample
Distribution by Regions
(Source: Test Publisher's Manual.)

	North-east	Mid-West	South-east	South-west	West
Total Population %	26.7	29.1	18.8	10.0	15.4
Weighted Sample Percent	19.04	34.71	24.11	13.85	8.29

4. Evidence of Ceiling and Floor Effects

   a. No students in the 9-11 grades received a perfect or near-perfect score on any of the subtests. One percent of the 12th graders received a near-perfect score on the composition test, while 2 percent of the 12th graders received a near-perfect score on the math test.

   b. The percentage of students (across the four grades) scoring at or below chance level on the reading, composition, and literature tests ranged from 1 percent to 8 percent; the mean was 4 percent. For the math test, the range was 16 percent of 22 percent.

5. Reliability Data

   a. Test-retest Reliability: No data available.

   b. Internal Consistency Reliability: Using a split-test procedure, the correlations ranged between .86 and .93 on the individual sub-tests. The composite scores (made up of subtest scores) correlated between .96 and .97 across the four grades.

   c. Alternate Form Reliability: No data available.

6. Validity Data

   a. Content Validity: Panels of subject-matter experts reviewed textbooks and other curriculum materials to identify current instructional objectives around which to develop and select test items. Subsequent to item development, pilot studies were conducted to identify appropriate test formats for each education level. Once tests were constructed, test tryouts were conducted to collect data that were used to drop, revise, and add items to the final tests.

b. Concurrent Validity: Intercorrelations among the Test of Academic Progress, Form S standard scores and subtest scores of the Cognitive Ability Test are presented in the technical reports. Generally, the correlation coefficients ranged between .70 and .80 for tests that were designed to measure the same or similar skills and abilities.

7. Items Debiasing Procedures Followed

There is no indication in the publisher's technical documents that item debiasing procedures were used.

Name of Test: Sequential Tests of Educational Progress (STEP Series II)

Publisher: Addison-Wesley Publishing Co., Inc. (formerly published by Educational Testing Service)

Test Levels: Forms 3A and 3B: Grades 7 - 9
Forms 2A and 2B: Grades 10-12
Forms 1A and 1B: Grades 13-14

1. Test Content, Number of Items, and Completion Time by Test Form

	Form 3A-B		Form 2A-B		Form 1A-B	
	Number of Items	Time in Minutes	Number of Items	Time in Minutes	Number of Items	Time in Minutes
a. Reading (1 score)						
Part I: Straightforward Comprehension and inference	30	15	30	15	30	15
Part II: Comprehension, translation, inference, analysis	30	30	30	30	30	30
b. Mechanics of Writing (3 scores)						
Spelling	45	15	45	15	--	--
Capitalization and Punctuation	45	25	45	25	--	--
c. English Expression (1 score)						
Part I: Grammar	40	20	40	20	40	20
Part II: Syntax	20	20	25	20	25	20
d. Mathematics Computation (1 score)						
Addition, subtraction, multiplication, division, elementary algebraic manipulations	60	40	60	40	--	--

	Form 3A-B		Form 2A-B		Form 1A-B	
	Number of Items	Time in Minutes	Number of Items	Time in Minutes	Number of Items	Time in Minutes
e. Mathematics Basic Concepts (1 score) Concepts include number and operations, measurement and geometry, relations, functions, graphs, proofs, probability and statistics	50	40	50	40	50	40

2. Norm Data Available For:

Forms 3A and 3B, Grades 7 - 9
Forms 2A and 2B, Grades 10-12
Forms 1A and 1B, Grades 13-14 (except for mathematics computation and mechanics of writing)

College freshmen norm data were collected in the fall of 1969; norm data for other grades (i.e., grades 3-12, and grade 14) were collected in the spring of 1970.

3. Representativeness of Standardization Samples

Numbers and Percentages of Public School Students in the National Population and in the Norms Sample (by Size of School System Enrollment) (Source: Test Publisher's Manual.)

Size of School System Enrollment	Students in Population		Students in Norms Sample	
	Number	Percent	Number	Percent
100,000+	5,604,629	13	7,4[illegible]	8
50,000-99,999	3,800,079	9	6.[illegible]	7
25,000-49,999	2,913,655	7	6,175	7
12,000-24,999	5,854,193	14	16,169	17
6,000-11,999	7,168,942	17	11,745	13
3,000-5,999	7,015,908	16	17,151	19
1,200-2,999	6,609,656	15	17,336	19
600-1,999	2,331,697	5	6,233	7
200-599	1,799,411	4	3,007	3
Total	43,098,200	100	91,507	100

Percentages of Elementary and Secondary School Students in the National Population* and in the Norms Sample (by Regions) (Source: Test Publisher's Manual.)

Grade	Region I		Region II		Region III		Region IV	
	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.
3	24	22	27	27	24	26	25	25
4	24	22	26	27	25	26	25	25
5	25	22	27	27	24	26	24	25
6	24	22	28	26	24	26	24	26
7	23	22	28	26	25	26	24	26
8	24	22	28	27	24	26	24	25
9	24	23	28	28	25	24	23	25
10	23	24	28	28	26	23	23	25
11	23	24	28	29	25	22	24	25
12	20	24	29	29	27	22	24	25
Catholic**	43	42	32	37	11	08	14	13

*Population values compiled from Statistics of State School Systems: 1965-1966. Regions defined in Statistics of State School Systems: 1961-1962.
Regions:
I - Conn., Del., D.C., Me., Md., Mass., N.H., N.J., N.Y., Pa., R.I., Vt.
II - Ill., Ind., Ia., Kan., Mich., Minn., Mo., Neb., N.Dak., Oh., S. Dak., Wisc.
III - Ala., Ark., Fla., Fa., Ky., La., Miss., N.C., S.C., Tenn., Va., W.Va.
IV - Ak., Ariz., Calif., Colo., Hi., Id., Mont., Nv., N.M., Ok., Ore., Tex., Ut., Wash., Wy.

**These percentages are based on enrollment data from The Official Catholic Directory for all grades combined since enrollments by grade were not available.

Numbers and Percentages of College Freshmen in the National Population and in the Norms Sample (Source: Test Publisher's Manual.)

Size of College Enrollment	Freshmen in Population*		Freshmen in Norms Sample	
	Number	Percent	Number	Percent
15,000+	61,705	7	25	2
10,000-14,999	83,009	10	77	5
5,000-9,999	163,781	20	222	16
3,000-4,999	154,979	19	280	20
1,500-2,999	154,624	19	348	25
1,000-1,499	87,008	11	177	13
500-999	84,786	10	165	12
1-499	32,088	4	96	7
Total	821,980	100	1,390	100

*Estimated from total enrollment data presented in Education Directory, 1968, Part 3, and opening day fall enrollment in Higher Education, 1968, Part B.

4. Evidence of Ceiling and Floor Effects

   a. Typically, about 1 percent of the students at each grade level received a perfect or near-perfect score on each test.

   b. For the reading, writing, and English tests, the range of students at each grade level scoring below the chance level was 1 percent to 4 percent. The range for the mathematics tests was 1 percent to 5 percent.

5. Reliability Data

   a. Test-Retest Data: not available.

   b. Internal Consistency Reliability: The KR20 coefficients range between .79 to .95. Most of the KR20 coefficients were at or above .88.

   c. Alternate Form Reliability: The correlation coefficients ranged between .58 and .93 across grades 8-14. However, a majority of the coefficients were .87 or above.

   d. Inter-Level Reliability: Publisher presents data on the relationships of raw scores to converted scores, and reports the following results:

- Generally, the forms at different levels are similar in difficulty.
- The exceptions are:
  - Forms 2A-3B of Mechanics of Writing are slightly more difficult than their counterparts, 2B and 3A.
  - Forms 2A and 4B of Spelling (writing test) are slightly more difficult than 3B and 4A.
  - Forms 3B and 4A of Capitalization and Punctuation (writing test) are slightly more difficult than 3A and 4B.
  - Forms 1B and 2A of Mathematics Basic Concepts are slightly more difficult than forms 1A and 2B.
- Publisher presents correlation data that resulted from calibrating adjacent level forms; these correlations range from .62 to .91 (the vast majority are within the .80 to .90 range).

6. Validity Data

a. Content Validity: Subject-matter specialists at ETS worked with committees of subject-matter experts (external consultants) to devise test specifications and items. A list of external consultants is included in technical Handbook. Sufficient quantity of items in each subject-matter area were developed to permit three or four pretests for each pair of the intended final forms. Consultants were used, along with ETS staff, to make final selection of items to be included in each test.

b. Concurrent Validity: Students were administered each test within the STEP Series II battery and each test within the SCAT series II battery. The SCAT is an ability test that includes verbal and quantitative sections. The SCAT verbal, quantitative, and total scores were correlated with each subtest score within the STEP Series II battery. These coefficients ranged between .54 and .84; a majority of the coefficients were in the .65 to .75 range.

7. Item Debiasing Procedures Followed

There is no indication in the Technical Report that item debiasing procedures were used.

Name of Test: California Achievement Tests, 1970 Edition (CAT-70)

Publisher: CTB/McGraw-Hill

Test Levels: Level 4, Grades 6-9
Level 5, Grades 9-12

1. Test Content, Number of Items, and Completion Time by Test*

	Number of Items	Time in Minutes
a. Reading		
Vocabulary	40	10
Comprehension	45	40
Totals	85	50
b. Math		
Computation	48	33
Concepts and Problems	50	23
Totals	98	56
c. Language		
Mechanics (capitalization and punctuation)	80	25
Usage and Structure	54	14
Spelling	32	8
Totals	166	47

2. Norm Data

Available for grades 9-12. Norm data were collected in February and March, 1970.

3. Representativeness of Standardization Sample

The standardization sample contained 203,684 students in Grades 1-12 from public and Catholic schools in 36 states. The public school sample was selected by following a stratified random sampling procedure that included the

* The data reported here are for the Level 5 test. The Level 4 test varies only slightly in the time required to complete it and the number of items used for each subtest.

following dimensions: (1) geographic region, (2) average enrollment per grade and (3) community type. A total of 355 public schools participated in the standardization program. Of this total, 349 schools completed and returned a School Characteristics Questionnaire that produced some socio-demographic data on the types of students that participated in the sample. However, none of the CTB/McGraw-Hill technical reports compared sample characteristics with the national population. See Table A-1 for additional sample data.

Data on minority group participation in the standardization sample was obtained by asking the following question: "About what percentage of the students in this school are white?" Results are as follows:

- Overall average response was 83 percent
- Average responses by size of districts were:
  - small districts: 90 percent
  - medium districts: 82 percent
  - large districts: 66 percent
- Average responses by type of community were:
  - urban: 67 percent
  - rural, town and other: between 87 percent and 88 percent

4. Evidence of Ceiling and Floor Effects

   a. Typically 1 percent to 2 percent of the students at each grade level received a perfect or nearly-perfect score on each of the subtests. The only exception was in the 12th grade; 4 percent of the students received a near-perfect score on the mathematics test.

   b. For the 9th and 10th grades, 2 percent to 5 percent of the students at each grade level scored below chance level on each subtest. Only 1 percent of the students scored similarly in grades 11 and 12.

Table A-1.

Number of Schools in Neighborhood Types According to Responses of School Characteristics Questionnaire

(Source: Test Publisher's Manual.)

Neighborhood Type	Elementary Schools Only	Secondary Schools Only	Both Elementary and Secondary	Total
Rural area	31	11	19	61
Residential suburb	46	29	0	75
Industrial suburb	7	3	0	10
Small town (5,000 or less)	21	14	5	40
City of 5,000 to 50,000	28	21	1	50
Residential area of a large city(50,000+)	21	9	0	30
Inner part of a large city (50,000 +)	14	5	0	19
Rural area and small town	23	7	4	34
Other	18	11	1	30
Total	209	110	30	349

5. Reliability Data

a. Test-retest Reliability: Not available.

b. Internal Consistency Reliability: The KB Formula coefficients ranged between .67 and .96 on the subtests for reading, mathematics, and language across grades 9-12. The vast majority concentrated between .85 and .95. For the total battery, the coefficients ranged between .97 and .98 for each subject area across grades 9-12.

c. Alternate Form Reliability: Alternate Form Reliability data are reported for grade 10 only. Product-moment coefficients ranged between .80 and .96. The average was .86.

d. Inter-Level Reliability: Inter-Level Reliability data are reported for grade 9 only. Product-moment coefficients ranged between .66 and .94. The average was .83

6. Validity Data

a. Content Validity: CTB/McGraw-Hill test development staff reviewed reading, mathematics, and language texts, among other course materials, to identify curricular objectives to be measured and textbook formats that typically are encountered by students. Test items were obtained from three sources: (1) newly written items, (2) items from the 1957 edition of the CAT that were updated and rewritten, (3) items from 1957 edition that were used intact. A number of tryouts were undertaken to collect item data that were used to delete, modify, or add items that made up the standardization battery. There is no evidence that external subject-matter specialists served as consultants during test revision activities.

b. Concurrent Validity: Not available.

7. Item Debiasing Procedures Followed

There is no indication that the CAT-70 was exposed to item debiasing procedures.

Name of Test: Comprehensive Tests of Basic Skills, Level 4, Form S and T (CTBS-4)

Publisher: CTB/McGraw-Hill

Test Level: Level 4, Forms S and T, Grades 8-12

1. Test Content, Number of Items, and Completion Time by Test

	Number of Items	Time in Minutes
a. Reading		
Test 1: Vocabulary	40	11
Test 2: Comprehension	45	35
Totals	85	46
b. Language		
Test 3: Spelling	30	12
Test 4: Language Mechanics	20	10
Test 5: Expression	35	27
Totals	85	49
c. Mathematics		
Test 6: Computation	48	40
Test 7: Concepts and Applications	50	30
Totals	98	70
d. Reference Skills		
Test 8: Reference Skills	20	10

2. Norm Data Available For: Level 4: Grades 8-12

3. Representativeness of Standardization Sample

To obtain background data on the standardization sample, a questionnaire was sent to each of the 285 participating schools in the spring of 1973. All 285 schools responded to the questionnaire. The data reported in the tables below were obtained from the questionnaire.

Schools Responding to the Questionnaire by Neighborhood Self-Description

(Source: Test Publisher's Manual.)

Neighborhood Self-Description	Elementary	Secondary	Both	Total
Rural Area	29	13	2	44
Residential Suburb	36	8	0	44
Industrial Suburb	2	1	0	3
Small Town	22	8	4	34
City of 5,000 to 50,000	49	15	0	64
Residential Area of a Large City	45	10	0	55
Inner Part of a Large City	37	4	0	41
Total	220	59	6	285

Percentage of Parents in Each Occupational Category by Neighborhood Type

(Source: Test Publisher's Manual.)

Occupation	Rural	Residential Suburb	Small Town	City	Large-City Residential	Inner City
Professional	7.0	14.5	7.3	10.4	12.8	4.2
White Collar	16.0	29.6	24.7	22.3	29.5	15.3
Skilled	42.5	43.0	45.9	37.9	37.3	30.6
Unskilled	34.4	13.9	22.7	28.2	18.4	50.4

School Ethnic Composition by Region
(in Percent)

(Source: Test Publisher's Manual.)

Ethnic Group	New England	Mid-east	Great Lakes	Plains	South-east	South-west	West
Black	11.8	19.2	16.5	9.5	31.1	8.7	10.0
Spanish-speaking	6.3	0.8	2.7	0.3	0.8	20.8	19.0
Other	82.0	80.0	76.7	90.2	68.1	69.1	71.1

School Ethnic Composition by Neighborhood Type
(in Percent)
(Source: Test Publisher's Manual.)

Ethnic Group	Rural	Residential Suburb	Small Town	City	Large-City Residential	Inner City
Black	6.4	3.8	3.2	21.2	19.5	43.4
Spanish-speaking	2.4	4.7	4.5	9.2	9.7	15.6
Other	88.7	91.5	92.3	68.8	69.2	41.0

4. Evidence of Ceiling and Floor Effects

   a. For the reading, language, mathematics, and reference skills tests, approximately 1 percent of the students at each grade level (8-10) received a perfect or near-perfect score. For grades 11 and 12, the range on the reading, language, and mathematics tests was 1 percent to 8 percent. For the reference skills test, 13 percent of the 11th graders and 21 percent of the 12th graders received a near-perfect score.

   b. For the 8-10 grades, the range of students scoring below chance level on each test was 2 percent to 18 percent; the mean was 9 percent. For grades 11 and 12, the range was 1 percent to 4 percent.

5. Reliability Data

   a. Test-Retest Reliability: For grade 10, test-retest reliability product-moment coefficients ranged between .59 and .91 across the subtests. The average was .82.

   b. Internal Consistency Reliability: The KR Formula 2[illegible] coefficients ranged between .76 and .98 for reading, mathematics, and language across grades 8-12. The majority of coefficients were in the .85-.93 range.

   c. Alternate Form Reliability: For grade 10, alternate form reliability product-moment coefficients ranged between .61 and .91. The average across subtests was .78.

d. Inter-Level Reliability: A sample of students in grade 8 was administered the Level 3 and Level 4 tests in order to evaluate inter-level reliability. The product-moment coefficients ranged between .69 and .94. The average coefficient was .81.

6. Validity Data

a. Content Validity: This test was developed from a careful and critical review of the best items in CTBS/Form Q. Revised items as well as new items were prepared by content specialists (external consultants) and CTB/McGraw-Hill staff. These were administered to a nationwide sample in the spring of 1972. Item data were reviewed and these served to make additional item revisions before standardization was undertaken.

b. Concurrent Validity: Not available.

7. Item Debiasing Procedures Followed

Prior to standardization, publishers used Black and Spanish-speaking reviewers who were teachers and content and curriculum experts to identify items that seemed to be biased against minorities. Such items were either revised or deleted from the test. In addition, statistical analyses were conducted to identify biased items, which were subsequently deleted.

Name of Test: California Achievement Tests, Form C and Form D (CAT-77)

Publisher: CTB/McGraw-Hill

Test Levels: Level 18, Grades 7.5 - 9.9

Level 19, Grades 9.5 - 12.9

1. Test Content

Number of items, and completion time by test.

	Number of Items	Time in Minutes
a. Reading		
Vocabulary	30	16
Comprehension	40	38
b. Language Arts		
Spelling	20	10
Language Mechanics	25	15
Language Expression	38	27
c. Mathematics		
Computation	40	28
Concepts and Applications	45	38

2. Norm Data

Norm data are available for grades 9-12, and were collected on the same students in the fall and spring.

3. Representativeness of Standardization Sample

The sampling strategy used for selecting the norming sample provided for proportionate representation of minority groups. The minority group designations used were Black, Spanish-speaking, or other.

4. Evidence of Ceiling and Floor Effects

a. Typically, less than 2 percent of the students at each grade level received a perfect or nearly perfect score on each of the subtests.

b. Less than 3 percent of the students at each grade level scored below chance level on each subtest.

5. Reliability Data: Not available*

6. Validity Data:

a. Content Validity: Current curriculum guides and other instructional materials available from state departments of education and most major cities were reviewed to identify the basic skills common to most curricula. Using the identified objectives, CTB/McGraw Hill item writers prepared items for the Tryout Edition of CAT-77. Panels of experts representing various ethnic groups reviewed and assisted with the elimination of seemingly biased items. With the Tryout Edition data collected, items analyses were performed to help modify the final version of CAT-77.

b. Concurrent Validity: Not available*

7. Item Debiasing Procedures Followed

Two procedures were used to reduce the possible presence of racial, ethnic, sex bias in the battery. Consultants from various ethnic groups reviewed the Tryout Edition of the CAT-77. Their suggestions were used to modify or eliminate material that appeared to be biased. The second procedure was to statistically analyze item data and to eliminate, where possible, items that seemed to be biased against particular ethnic groups.

---

*At the time of this review, SDC did not have available the reliability and validity data on the CAT-77.

APPENDIX B

STUDY SKILLS TEST SUMMARIES

Name of Test: Effective Study Test (EST)

Author: William F. Brown

Publisher: Effective Study Materials, San Marcos, Texas

Test Levels: Form H - Grades 9-12
Form C - College Freshmen and Sophomores

1. Test Content

A. Cognitive Content

Scales	# Items/Scale
a. Reality Orientation: measures realistic understanding of the problems connected with developing effective study habits.	25
b. Study Organization: measures knowledge about effective methods for budgeting study time and organizing study area.	25
c. Writing Behavior: measures knowledge about effective methods for taking class notes and writing themes and reports.	25
d. Reading Behavior: measures knowledge about effective methods for reading textbooks and rembering the material.	25
e. Examination Behavior: measures knowledge about effective methods for preparing for and taking objective and essay tests.	25
Total Study Effectiveness: combines scores on all five scales to provide a single overall measure of knowledge about effective study and methods and the factors influencing their development.	
TOTAL:	125 items

B. Study Habits Content

The test assesses a student's knowledge about efficient study skills rather than assessing the student's actual study behavior.

2. Grade Appropriateness

Form H - Grades 9-12

Form C - College students (Freshmen and Sophomores)

The reading level appears very high and might be difficult for some students tested.

3. Norm Sample

Norms for the EST were derived from test scores for 2,354 high school students and 2,047 college freshmen tested during the 1962-63 school year. The normative sample includes all pupils in grades 9-12 at six Texas high schools, as well as all entering freshmen at three Texas colleges. The norm sample does not include college sophomores.

4. Ceiling and Floor Effects

Less than one percent of the students (at each grade level) received a perfect or near-perfect score on the total test battery.

b. About one percent of the students scored at or below the chance level.

5. Reliability Data

a. Test-Retest: Not available

b. Internal Consistency: Not available.

c. Alternate Form Reliability: Not available.

d. Inter-Level Reliability: Not available.

6. Validity Data

a. Content Validity

Form C - College Level EST: The development of this form included extensive review of how-to-study literature and other tests designed to measure study skills and habits. The entire development process took about five years to complete and required four major revisions in the items used in the EST. During the development, several field tryouts were conducted and the data were used to conduct item discrimination and item difficulty analyses. Appropriate item revisions were made as a result of these analyses. There is no indication in the test manual that external consultants were used in the development of the college level EST.

Form H - High School Level EST: This form was developed by revising the college level form. Assisting the senior author were committees of high school teachers and counselors that revised Form C to conform to high school instructional procedures, academic requirements, and study conditions.

b. Predictive Validity

The following tables give the validity coefficients of the EST with GPA and with an ability to test at the college and high school levels.

Table 1.

Correlations Between One-Year Grade Averages, Total EST Scores, and Total SCAT Scores (Source: Test Publisher's Manual.)

College	N	Correlation Coefficients*	
		EST with GPA	EST with SCAT
Southwest Texas State College	238	.54	.28
Texas Lutheran College	876	.57	.29

*EST = Effective Study Test total score; GPA = first-semester grade-point average; SCAT = School and College Ability Test total score.

Table 2.

Correlation Between One-Year Grade Averages and Effective Study Test Total Scores for 2354 Students in Grades 9-12 at Six Texas High Schools (Source: Test Publisher's Manual.)

HIGH SCHOOL	GRADE 9		GRADE 10		GRADE 11		GRADE 12		TOTAL GRADES*	
	N	r	N	r	N	r	N	r	N	r
San Marcos	267	.57	235	.58	181	.64	146	.60	829	.59
Calhoun	241	.55	232	.62	137	.54	100	.61	710	.58
Canyon	94	.56	80	.55	65	.65	41	.54	280	.58
Navarro	28	.56	21	.52	15	.95	20	.61	84	.68
Marion	33	.59	32	.70	36	.74	26	.59	127	.67
Randolph Field	122	.77	107	.55	63	.57	32	.52	324	.65
TOTAL SCHOOLS:	785	.61	707	.59	497	.63	365	.59	2354	.60

*Fisher's Z-function was used to obtain the weighted averages of obtained correlations.

c. Concurrent Validity

The following table gives the validity coefficients of the EST and the SSHA at the college and high school levels.

Table 3.

Concurrent Validity Data for College and High School Freshmen (Source: Test Publisher's Manual.)

Colleges	N	Correlation Coefficients*
		EST with SSHA
Southwest Texas State College	238	.34
Texas Lutheran College	876	.36
High School		
San Marcos High School	207	.43

*EST = Effective Study Test total score; SSHA = Survey of Study Habits and Attitudes.

7. Test Administration Characteristics

The EST is administered in a group setting by the teacher/proctor. No special training is required to administer the test. The students complete a "true/false" answer sheet that is hand-scored using an answer key. A special answer sheet would have to be provided if machine scoring is to be used. The test requires between 40-45 minutes to complete.

Name of Test: Survey of Study Habits and Attitudes (SSHA)

Publisher: The Psychological Corporation, New York

Test Levels: Form H - Grades 7-12
Form C - Grades 12-14

1. Test Content

A. Cognitive Content

The survey is not designed to be a cognitive test.

B. Study Habits Content

The survey is designed to measure study habit strengths and weaknesses by asking the students to agree or disagree with statements about ways of studying and about attitudes on education, teachers and school activities.

It has four basic scales--two subtotals and a total score. The scoring keys are the same for both forms; however, the wording of the items differs.

Scale	# of Items
a. Delay Avoidance (DA): promptness in completing assignments, lack of procrastination, freedom from delay and distraction.	25
b. Work Methods (WM): use of effective study procedures, efficiency in doing academic assignments, and how-to-study skills.	25
c. Study Habits (SH): combines the scores on DA and on WM to provide a measure of academic behavior.	(Subtotal) 50

Scale (Cont'd)	# of Items
d. Teacher Approval (TA): opinions on teachers and their classroom behaviors and methods.	25
e. Education Acceptance (EA): approval of educational objectives, practices, and requirements.	25
f. Study Attitudes (SA): combines the scores on the TA and EA scales to provide a measure of scholastic beliefs.	(Subtotal) 50
g. Study Orientation (SO): combines scores on SH and SA scales to provide an overall measure of study habits and attitudes.	
GRAND TOTAL:	100

2. Grade Appropriateness

The SSHA is designed in two forms for grades 7 through 14. However, the vocabulary level seems a bit difficult for some high school students.

3. Norm Sample

At the college level, the SSHA was normed in 1964 using 3,054 college freshmen at nine colleges: Antioch, Bowling Green State, Colorado State, Reed College, San Francisco State, Southwest Texas State, Stephen F. Austin State, Swarthmore College and Texas Lutheran College. College sophomores were not included in the norm sample. There is no evidence that this sample is nationally representative. At the high school level, 11,218 students in grades 7-12 were used to norm Form H. Sixteen high schools were used and they were located in Texas (10 schools); Colorado (2 schools); Illinois (1 school); Maryland (1 school); Utah (1 school); and Missouri (1 school). Most of these high schools are located in small towns, thus throwing the sample's representative's into question.

4. Ceiling and Floor Effects

a. Between 1% and 3% of the students received a perfect or near-perfect score on the Form C or on the Form H test.

b. The percentage of college students that scored at or below the chance level ranged from 1% (Study Orientation Total Score) to 10% (Delay Avoidance subscale, and the Study Habits subscale). The percentage of high school students that scored at or below chance level ranged from 3% (Study Orientation Total Score, and the Study Attitudes subscale) to 15% (Delay Avoidance subscale).

5. Reliability Data

Form C - College Level

a. Test-Retest Data: Two studies were conducted, one on a sample of 144 freshmen with a four-week interval between administrations, and the other on a sample of 51 freshmen with a fourteen-week interval. The test-retest coefficients for the four-week interval were .93, .91, .88., and .90, respectively for the four subscales (DA, WM, TA, and EA). The corresponding coefficients for the fourteen-week interval were .88, .86, .83, and .85, respectively.

b. Internal Consistency: For the 465 freshmen tested at Southwest Texas State in the fall of 1960, reliability coefficients for the four subscales ranged from .87 to .89.

c. Alternate Form Reliability: Not available

d. Inter-Level Reliability: Not available

Form H - High School Level

a. Test-Retest Data: 237 ninth graders were tested in Texas with a four-week interval between sessions. The test-retest reliability coefficients were .95, .93, .93, and .94, respectively for the four subscales, and .95 for the SSHA total score.

b. Internal Consistency: There is no evidence in the available materials of an internal consistency test at the high school level.

c. Alternate Form Reliability: Not available

d. Inter-Level Reliability: Not available

6. Validity Data

a. Content Validity

Form C - College Level

- Form C of the SSHA was developed over a period of eight years, and was subjected to numerous field tests and revisions.

- An interim revision involved the sorting and placing of items into subscales in accordance with the subjective judgment of a 15-member panel of psychologists. The content of the resulting subscales was agreed upon by 80 percent of the judges.

- The final revision involved a redistribution of items based on the advice of counselor trainees who had used the SSHA.

- The revised Form C was then evaluated statistically and it was concluded that each of the subscales measured traits that played a significant role in academic achievement.

Form H - High School Level

- Form H of the SSHA was developed by revising Form C in accordance with junior and senior high school instructional procedures, academic requirements, and study conditions.

- The revisions were accomplished by committees of teachers and students working under the direction of the senior author.

b. Predictive Validity

The following table gives the validity coefficients of SSHA with GPA and aptitude tests at the college level (Form C).

Table B-1.

Correlation of SSHA--Form C with a Scholastic Aptitude Test[a] and with Grade Point Average,[b] and Multiple Correlation of Scores with Grade Point Average (Source: Test Publisher's Manual.)

College	N	Correlation Coefficients			$R_{3.12}$	Means and Standard Deviations					
		SSHA(1) with GPA(3)	SSHA(1) with Aptitude Test(2)	Aptitude Test(2) with GPA(3)		SSHA(SO) Total Score Mean	SD	Aptitude Test Mean	SD	GPA Mean	SD
1	198	.35	.20	.48	.55	116.1	27.8	107.2	20.0	2.2	.8
2	379	.25	.07	.27	.36	125.5	27.2	312.6	12.6	2.2	.8
3	127	.33	.05	.23	.39	120.3	26.1	1272.7	116.8	1.7	.3
4	465	.45	.27	.59	.66	113.4	31.7	57.4	15.2	2.0	.8
5	372	.37	.25	.42	.50	110.5	29.1	303.1	14.6	1.9	.9
6	231	.40	.27	.51	.58	107.0	28.8	298.1	14.9	1.9	1.0
Weighted Average		.36	.21	.45	.53						

[a] Scholastic aptitude test: College #1 used American Council on Education Psychological Examination; College #3, CEEB-SAT; all other schools used School and College Ability Test (SCAT).

[b] Grade point average for first semester based on scale in which A=4, B=3, C=2, D=1, F=0.

[c] Obtained by converting each r to its Fisher's Z-function, weighting by the appropriate number of cases, averaging the values, then reconverting.

The following is a summary of the validity coefficients using GPA as criterion at the high school level (Form H):

Scale	Range of Coefficients	Weighted Averages
Delay Avoidance	.20-.56	.41
Work Methods	.31-.68	.47
Teacher Approval	.14-.53	.35
Education Acceptance	.25-.63	.48

c. Concurrent Validity

The manual cites as concurrent validity the results of a study comparing teacher academic subject ratings of students at one high school in Michigan and SSHA Scale Scores (see Table B-2).

Table B-2.

Percent of Students in Each Academic Rating Group[a] Scoring in Each Percentile Group on Three Scales of SSHA--Form H, By Grade (Source: Test Publisher's Manual.)

Percentile Group on SSHA Scale	Percent Receiving Each Rating								
	Study Habits			Study Attitudes			Study Orientation		
	A-B	C	D-E	A-B	C	D-E	A-B	C	D-E
Grade 7 (N=144)									
76-100	35	21	12	29	16	4	21	17	8
26-75	53	55	19	53	58	50	59	58	23
0-25	12	24	69	18	26	46	12	25	69
Grade 8 (N=130)									
76-100	28	12	6	35	9	6	36	9	0
26-75	56	55	41	50	52	47	49	52	59
0-25	16	33	53	15	39	47	15	39	41
Grade 9 (N=99)									
76-100	24	25	5	27	22	5	27	22	5
26-75	57	50	52	57	52	57	57	53	48
0-25	19	25	43	16	26	38	16	25	47

[a]Ratings were based on academic performance in English, social studies, mathematics, and science. Teachers assigned students to one of three categories: A-B, C, or D-E.

Other concurrent validity data were provided in the Effective Study Test evaluation summary (Table III).

7. Test Administration Characteristics

The SSHA is administered in a group setting by the teacher/proctor. No special training is required to administered the test. The students use an

IBM answer sheet that can be hand or machine scored. Students respond to a multiple choice format for each question (e.g., "rarely," "sometimes," "frequently," "generally," "almost always,"). Estimated completion time is 25-35 minutes.

Name of Test: The Study Skills Counseling Evaluation

Author: George Demos

Publisher: Western Psychological Services

Test Levels: Grades 10-12

Two-year and four-year colleges

1. Test Content

   A. Cognitive Content:

   The test is not designed to measure the student's cognitive knowledge.

   B. Study Habits Content:

   The test is designed to measure a student's strengths and weaknesses in several areas that pertain to study habits. These areas, along with the number of items used within each area, are as follows:

		Number of Items
a.	Study-time distribution	4
b.	Study conditions	2
c.	Taking notes	9
d.	Preparing and taking examinations	9
e.	Other habits and attitudes	9
f.	Critical items of special importance. (These items have been found to differentiate B and C students.)	17
		50 (Total test items)

Examples of "Critical Items..." are:

12. I study the same subject in the same place at the same time.
13. I study my notes at different times.
25. I put what I read or hear into my own words whenever I can.
35. I use in a practical way the course material I study.

2. Grade Appropriateness

There is one form of the SSCE for use at the secondary and postsecondary levels. The test is designed for use in grades 10-12 and the four years of college.

3. Norm Sample

The test manual (provided by WPS) indicates that "students in the norm group were obtained by randomly sampling students in introductory classes at several Los Angeles area schools." Separate norm data are available for students in high school, a two-year, and four-year college. There is no evidence that the norming samples included proportionate minority student representation.

4. Ceiling and Floor Effects

a. Less than one percent of the students in each norming sample received a perfect or near-perfect score on the total battery.

b. About one percent of the students in each sample scored at or below the chance level.

5. Reliability Data

a. Test-Retest: A test-retest study, using a sample of 74 students from California State University-Northridge, produced a reliability coefficient of .94. There was a one-week interval between each testing.

b. Internal Consistency: Not available.

c. Alternate Form: Not appropriate.

d. Inter-Level: Not appropriate.

6. Validity Data

a. Content Validity: The publisher's manual does not describe how the SSCE was developed.

b. Predictive Validity: The SSCE was validated on 172 students attending Los Angeles Valley College. Midterm grade point average in academic subjects was the criterion measure used. The coefficient was .38 and statistically significant.

c. Concurrent Validity: The SSCE performance data were also correlated with data from the Wrenn Study Habits Inventory; the correlation coefficient was .31 (no comment was made about statistical significance).

7. Test Administration Characteristics

The SSCE can be group administered by a teacher and requires about 20 minutes to complete. The results can be machine scored or hand scored.

APPENDIX C

AFFECTIVE TEST SUMMARIES

Name of Test: Self Appraisal Inventory (SAI)

Publisher: Instructional Objectives Exchange, Los Angeles, California.

Test Level: Secondary - Grades 7-12

1. Test Content

The inventory contains 62 statements with which the student can agree or disagree on a 4-point scale (e.g., strongly agree, agree, disagree, strongly disagree). The statements pertain to four dimensions of self-concept. Although the dimensions may be scored separately, items relating to each dimension are interspersed rather than grouped.

Dimension	# of Items
a. Peer: indicates positive peer relationships or perceptions of the self in social situations	16
Item Examples: 3. I ought to get along better with other people. 13. I am well liked by kids my own age.	
b. Family: perceptions of self in terms of family situations or relationships	16
Item Examples: 10. Usually no one pays much attention to me at home. 36. I feel free to argue with my family.	
c. School: perceived success or failure in scholastic endeavors	14
Item Examples: 11. Getting good grades is pretty important to me. 32. I am slow in finishing my school work.	
d. General: perception of general self esteem	16
Item Examples: 2. I am satisfied to be just what I am. 17. I am often not a happy person.	
TOTAL	62

2. Level Appropriateness

The inventory is designed for students from grades 7 through 12. The vocabulary level seems appropriate for grades 7 through 12.

3. Floor and Ceiling Effects

There is no information on the distribution of test scores.

4. Reliability Data

a. Test-retest: see table below.

b. Internal Consistency: see table below.

Internal Consistency and Stability Estimates for The Revised Measures
(Source: Measures of Self-Concept K-12, revised edition, I.O.X.)

Name of Measure	Internal Consistency Index (KR-20)		2 Week Test-retest Stability Index	
	n	r	n	r
Self Appraisal Inventory	87	.75	141	.87
General	133	.60	178	.67
Peer	128	.61	179	.62
Scholastic	101	.72	153	.53
Family	124	.74	185	.70

c. Alternate Form Reliability: Not available.

d. Inter-Level Reliability: Not available.

5. Validity Data

a. Content Validity: The test manual states that throughout the various stages of revision of the test major consideration was given to the content validity. The developers periodically discarded certain items. No quantified indices of validity are given, and external experts apparently were not used to review items.

b. Concurrent Validity: Not available.

6. Sample Characteristics

From the materials available, there is no indication that norming or standardization was done. Early test tryouts utilized students in Florida, Kansas, Missouri and California. The final revision involved ten schools in the Culver City, California, Unified system plus the Marina del Rey Jr. High School in Los Angeles, CA. Five of these schools were identified by their administrative staffs as representing low SES, five as representing middle SES, and one as representing high SES. 1,229 pupils were field tested.

7. Administration

The "Self Appraisal Inventory" is intended to be administered to a group by a teacher/proctor who reads directions and explains the use of the answer sheet. It is emphasized that there are no "right" or "wrong" answers and that, in order to obtain maximum participation, both perceived and real anonymity must be maintained. The student responds to a 4-point Likert-type scale on an answer sheet which may be either hand scored or arranged to be machine scored. There is no time limit; however, estimated completion time is 15-20 minutes.

Name of Test: For All I Know

Publisher: Instructional Objectives Exchange, Los Angeles, California

Test Level: Secondary - grades 7-12

1. Test Content

This inventory consists of ten hypothetical situations which describe backgrounds for achievement in school, scholastic integrity, confidence in school work and scholastic initiative. The alternatives present two positive and two negative behaviors in respect to these areas. Students will display positive self concepts by selecting positive alternatives.

Item Examples: 1. You have been given a class quiz. The instructions for one part of the quiz are not clear to you although the other students have already begun the test. You would probably

a. Call the teacher over to your desk and ask for help.
b. Begin to take the test as well as you could.
c. Ask the person sitting in front of you for help.
d. Complete all parts of the test except the one that is unclear to you.

3. One of your teachers has assigned an extra credit project. Students are not required to complete the assignment, but they may wish to increase their grades in the course by doing the project. So far, your grade is above average, so you would probably

a. Complete the project anyway.
b. Choose an interesting project since you don't have to worry about your grade.
c. Decide not to do the project.
d. Wait to find out how many of the good students are doing projects.

2. Level Appropriateness

The inventory is designed for students from grades 7 through 12. The vocabulary level seems appropriate for grades 7 through 12. With minor word changes, this test would be appropriate for college freshmen and sophomores.

3. Floor and Ceiling Effects

There is no information on the distribution of test scores.

4. Reliability Data

a. Test-retest: The materials show a two-week test-retest reliability of .31 using a sample of 184 students.

b. Internal Consistency: Using a sample of 135 students, the internal consistency reliability coefficient (KR20) was .74.

c. Alternate Form Reliability: Not available.

d. Inter-Level Reliability: Not available.

5. Validity Data

a. Content Validity: The manual states that throughout the various stages of revision of the test, major consideration was given to the content validity. The developers periodically discarded certain items. No quantified indices of validity are given, and external experts apparently were not used to review items.

b. Concurrent Validity: Not available.

6. Sample Characteristics

From the materials available, there is no indication that norming or standardization was done. Early test tryouts utilized students in Florida, Kansas, Missouri and California. The final revision involved ten schools in the Culver City, California, Unified system plus the Marina del Rey Junior High School in Los Angeles, California. Five of these schools were identified by their administrative staffs as representing low SES, five as representing middle SES, and one as representing high SES. 1,229 pupils were field tested.

7. Administration

The "For All I Know" inventory is intended to be administered to a group by a teacher/proctor who reads directions and explains the use of the answer sheets. It is emphasized that there are no "right" or "wrong" answers and that, in order to obtain maximum participation, both perceived and real anonymity must be maintained. The student responds to the hypothetical situations by marking on a separate answer sheet his choice of behavior in the situation. The answer sheets may be hand scored, or arranged to a machine scored. There is no time limit, however, estimated completion time is 10-15 minutes.

Name of Test: Self Concept of Ability Scale (SCAS)

Publisher: College of Education, Michigan State University, East Lansing, Michigan

Test Level: Form A: General

1. Test Content

Using eight multiple choice questions, the scale attempts to determine how a person feels about himself as a student. All eight items relate directly to the student's perceived ability to function in school.

Item Example: 1. How do you rate yourself in school ability compared to your close friends?

a. I am the best
b. I am above average
c. I am average
d. I am below average
e. I am the poorest

(Note: All answers are in the above sequence, from best to worst)

2. Sample Characteristics and Level Appropriateness

The scale was developed in 1962 using a sample of urban seventh grade students (513 males and 537 females). Black students were excluded from the analysis on the assumption that their self-concept of ability would differ from those of Anglo students. Although the scale developers make no recommendations about its grade appropriateness, the vocabulary level seems appropriate for grades 7 through 14. (However, for use at the college level minor word changes would have to be made.)

3. Floor and Ceiling Effects

There is no information on the distribution of the scores.

4. Reliability

The eight items form a Guttman scale with a reprodcibility coefficient of .95 for males and .96 for females. The scale has an internal reliability of .82 for males and .77 for females.

5. Validity

a. Content Validity: No data were available.

b. Concurrent Validity: No data were available.

Correlations Between Seventh Grade GPA, Intelligence, and Self-Concept of Ability[a]*

Variables Correlated	Correlation Coefficients: No Variables Controlled — Males N=513	Females N=537	Third Variable Controlled — Variable Controlled	Males N=513	Females N=537
7th GPA - IQ	.61	.65	S.C.	.48	.53
7th GPA - SC	.57	.57	IQ	.42	.39
Self Concept-IQ	.46	.48	GPA	.17	.17

*This table is from: W.B. Brookover, S. Thomas, and A. Paterson, "Self-Concept of Ability and School Achievement," Sociology of Education, Vol. 37 (1964), pages 271-278.

[a]The Multiple Correlation (R 1.23)of Self-Concept and I.Q. with GPA is .69 for males and .72 for females.

6. Administration

The available materials contain no information about the administration of the scale. However, the scale itself is organized in terms of eight questions with five possible answers to each and the student is directed to circle the response that best answers each question. It would appear to require a minimum of three minutes to complete. Maximum time would depend on the reading level of the respondent.

Name of Test: Attitude Toward School (this includes five separate instruments which are named below under Test Content)

Publisher: Instructional Objectives Exchange, Los Angeles, CA

Test Level: Secondary - Grades 7-12

Note: The Instructional Objectives Exchange has developed five independently scored measures of attitude toward school. Reliability and validity data were computed on the same sample at the same time. The five will be discussed separately only in terms of test content and administration procedures.

1. Test Content

A. School Sentiment Index (SSI)

This Index consists of 82 statements regarding various aspects of school, with which the student can agree or disagree on a 4-point scale (e.g., strongly agree, agree, disagree, strongly disagree). The statements pertain to five aspects of attitude toward school which may be scored separately or as a whole. Items relating to each aspect are interspersed rather than grouped.

Measurement Dimensions	# of Items
1) Teacher	
Mode of Instruction: adequacy, fairness, grading	15
Item Examples: 10. Most of my teachers try to make their subjects interesting to me.	
32. Most of the decisions in my classes are made by the teachers.	
Authority and Control: fairness, effectiveness, manner	10
Item Examples: 6. Most of my teachers do not allow students much choice in what they study in class.	
59. Most of my teachers are still fair with me as a person even when I've done poorly in my school work.	
Interpersonal Relationship with Students: consideration, friendliness and concern	12
Item Examples: 3. Most of my teachers seem interested in the things I do outside of school.	

Measurement Dimensions (Cont'd) # of Items

28. Most of my teachers don't try very hard to understand young people.

2) Social Structure and Climate: perceptions of the bureaucracy, school organization, traditions, and activities. 15

Item Examples: 24. I think there is too much pressure in school.
60. There are enough different groups here at school for any type of student to find friends.

3) Learning: interest and/or involvement in learning and related activities (e.g., homework, assignments, extra school work). 11

Item Examples: 11. I hate having to do my homework.
52. My favorite classes, regardless of subject, are those in which I learn the most.

4) Peer: friendliness, distance, stratification, peer group friendship patterns. 8

Item Examples: 9. Students here are not as friendly as in other schools.
41. I really like most of the kids at this school.

5) General: holding power of school 11

Item Examples: 4. Each morning I look forward to coming to school. TOTAL: 82
54. School is just a place to keep kids off the street.

B. Imagine That

This inventory seeks to measure a student's attitude regarding teacher behavior by presenting a series of 11 hypothetical situations. For each situation presented, the student selects from four alternatives the one which he feels is the most likely explanation for the described teacher's behavior. Two alternatives place the teacher in a favorable light and two view the teacher negatively. Specifically, the student may react to teacher behavior in the areas of mode of instruction (four items), authority and control (three items), and interpersonal relationships with pupils (four items).

Item Example: 1. You are sitting in a class in American History. Near the end of the hour of quiet study, the teacher mentions that she would like to speak to you after class. She probably wants to speak to you about:

a. A paper you wrote that she thought was particularly interesting.
b. Something you did in class that displeased her.
c. Giving you help on your current assignment.
d. Your poor work habits.

C. High School on T.V.

This measure presents students with 12 hypothetical situations, each accompanied by three alternatives; some of the alternatives present favorably and some unfavorably the school social structure and climate. Students select the alternative in each item which depicts the most realistic details, based on their own school experience, for inclusion in a television script which is to portray life at their school. Aspects of structure and climate include the administration, rule structure, openness of the administrative system, and availability of extra help.

Item Example: 4. Dick and Sue need to talk to their English teacher about an assignment. They go to her classroom to see her before class:

a. She is having coffee with another teacher and tells them to go ask their classmates.
b. She is busy grading papers and tells them they're old enough to think for themselves.
c. She is busy, but makes arrangements to see them another time.

D. What Would Happen?

The instrument consists of 11 fictitious situations involving two new students at school. Each situation is accompanied by four alternative consequences--two positive and two negative. The respondent is instructed to select the alternatives describing what would probably happen to the new students at their school. The students will exhibit positive attitudes toward peers by selecting positive alternatives.

Item Example: 1. This is Will's first day at school. He is looking for his classroom and stops a student in the hall to ask directions. The most probable response is that the other student:

a. Notices that Will is a new student, tells him where his room is, and asks him where he is from.
b. Ignores the question because he is in a hurry and races off for his own class.
c. Provides the room number and walks on.
d. Tells Will where the room is and asks him to have lunch with him and some of his friends.

E. Subject Area Preference (SAP)

This inventory is composed of a list of eight subject areas commonly taught in junior and senior high school. Each subject area is accompanied by six descriptive word scales (positive to negative, seven degrees) on which students mark their personal reactions to that subject area.

Item Example: Foreign Language

confusing	____:	____:	____:	____:	____:	____:	____:	clear
meaningful	____:	____:	____:	____:	____:	____:	____:	meaningless
good	____:	____:	____:	____:	____:	____:	____:	bad
complex	____:	____:	____:	____:	____:	____:	____:	simple
unimportant	____:	____:	____:	____:	____:	____:	____:	important
useful	____:	____:	____:	____:	____:	____:	____:	useless

The inventory provides an index of the students relative preference among given subjects. Subject areas included are: English, Math, Social Studies, Foreign Language, Art, Vocational Education, Science, and Physical Education.

2. Level Appropriateness

The instruments are designed for students from grades 7 through 12. It appears that the vocabulary level is appropriate for grades 7 through 12. Each of the separate subscales (except for "High School on T.V.") would be appropriate for the college level if minor word changes were made.

3. Floor and Ceiling Effects

There is no information on the distribution of test scores.

4. Reliability Data

a. Test-Retest: See table below.

b. Internal Consistency: See table below.

Internal Consistency and Stability Estimates
for the Revised Attitude Toward School Measures.
(Source: Attitude Toward School K-12, revised edition, I.O.X.)

Name of Measure	Internal Consistency Index (KR-20)		Two Week Test-Retest Stability Index	
	n	r	n	r
School Sentiment Index Total	47	.88	80	.49
Teacher: Instruction	74	.73	101	.68
Teacher: Authority	75	.73	107	.65
Teacher: Peers	76	.76	104	.81
Soc. Structure & Climate	78	.77	105	.64
Learning	74	.68	104	.62
Peers	73	.71	100	.71
General	72	.79	111	.58
Imagine That	90	.58	113	.51
High School on T.V.	83	.54	107	.61
What Would Happen?	86	.54	109	.54
Subject Area Preferences		(.94)*		(.97)*
English	90	.74	115	.78
Math	94	.73	117	.80
Social Studies	90	.62	115	.84
Foreign Language	88	.62	111	.84
Art	92	.59	116	.76
Vocational Education	90	.45	114	.80
Science	89	.67	114	.86
Physical Education	92	.71	116	.84

*These are reliability estimates for the total test. They were calculated by SDC using the Spearman-Brown Formula and Fisher Z scores. These coefficients were used to rate the SAP on the reliability criteria.

c. Alternative Form Reliability: Not available.

d. Inter-Level Reliability: Not available.

5. Validity Data

a. Content Validity: The study states that throughout the various stages of revision of the test, major consideration was given to the content validity. The developers periodically discarded certain items. No quantified items of validity are given, and external experts apparently were not used to review items.

b. Concurrent Validity: Not available.

6. Sample Characteristics

From the materials available, there is no indication that norming or standardized was done. Early test revisions utilized students in Florida, Kansas, Missouri, and California. The final revision involved ten schools in the Culver City, California, Unified system plus the Marina del Rey Junior High School in Los Angeles, California. Five of these schools were identified by their administrative staffs as representing low SES, five represented middle SES, and one as representing high SES. The number of pupils field tested was 1,229.

7. Administration

All five instruments are intended to be administered to a group by a teacher/proctor who reads directions and explains the use of the answer sheet. It is emphasized that there are no "right" or "wrong" answers, and that in order to obtain maximum participation, both perceived and real anonymity must be maintained. There are no time limits. Answer sheets may be hand-scored or arranged to be machine-scored.

Name of Test: Feelings About School (FAS)

Author: Robert E. Bills

Publisher: University of Alabama Press

Test Level: Grades 4-12 (2 forms - 1 and 2)

Note: The primary source of the following information was a telephone conversation with the author of FAS.

1. Test Content

The FAS consists of statements about school to which the student responds as being true or false from his point of view. The two forms (1 and 2) mirror each other and each includes 50 items. FAS 1 and 2 are essentially unidimensional measures of positive and negative attitudes toward teacher behavior, with a few items reflecting attitudes on "how the school is run."

Item Examples: 1. I am learning things in school which help me now. (True/False)
3. Too many of our teachers have "pets." (True/False)

2. Level Appropriateness

The FAS has been used in grades 4 through 12. To be appropriate beyond high school, the test would require a moderate amount of item revision.

3. Floor and Ceiling Effects

No frequency or percentile distribution of scores is available. The possible range of scores is -50 to +50. In field tests, there were very few perfectly positive, and no completely negative scores. The average was +10.5; however, actual scores varied widely according to sex and grade level. Highest scores occurred at grade 4 (average = +18), and lowest scores at grade 11 (average = +2).

4. Reliability Data

a. Test-Retest: Not available

b. Internal Consistency: Based on a sample of 2,691 students, the internal consistency coefficients of FAS 1 and 2 were .89 and .91, respectively.

c. Alternate Form: For 2,691 completed tests the scores for FAS 1 were correlated with the scores for FAS 2, and yielded a coefficient of equivalence of .90.

d. Inter-Level Reliability: Not available.

5. Validity Data

a. Content Validity: Dr. Robert E. Bills provided (personal communication) evidence of an extensive development effort to ensure the content validity of the FAS. The original scale was revised after various field tests and factor analyses.

b. Concurrent Validity: Not available.

6. Sample Characteristics

The sample of 2,691 students completing FAS forms included approximately the same number of males and females from grades 4 through 12. It was composed primarily of Anglo students in medium to high SES schools, but includes one low SES black high school. The schools were located in South Carolina; Kingston, Jamaica; Massachusetts; California; Florida; Missouri, and Alabama. There were more high school than elementary students and more Anglos than blacks.

(Note: The author is presently in the process of standardizing the FAS. The sample will represent various ethnic groups, both sexes, and grades 4-12 in a number of Southeastern schools primarily in Louisiana and Alabama.)

7. Administration

The FAS is intended to be administered to a group by a teacher/proctor who reads directions and explains the use of the answer sheet. Depending on the reading level of the group, the administrator may read items to the group or allow them to read the FAS themselves. Anonymity is stressed and it is emphasized that there are no "right" or "wrong" answers. The response format may be a hand-scored or machine scored. The FAS can be completed in 15 to 20 minutes.

Name of Test: Multidimensional I.E. Scale

Location: Gurin, P.; Gurin, G.; Lao, R.C., and Beattie, M. Internal-external control in the motivational dynamics of Negro Youth. Journal of Social Issues, 1969, 25, 29-53.

Test Level: High school and college

1. Test Content: This instrument is a modified form of Rotter's I.E. Scale. It consists of the 23 items from Rotter's scale--which pairs internal statements with external statements--3 items from the Institute for Social Research Personal Efficiency Scale, and 13 items written especially to tap beliefs about the operation of personal and external forces in the race situation in the United States. Four scoring factors have been developed:

	# of items
a. Factor I: Control Ideology--Measures respondent's beliefs about the role of internal and external forces in determining successes and failures of people in the culture at large. Item example: 1. (a) Without the right breaks, one cannot be an effective leader. (b) Capable people who fail to become leaders have not taken advantage of their opportunities.	13
b. Factor II: Personal Control--Measures respondent's beliefs in his/her own control. Item example: 14. (a) I have often found that what is going to happen will happen. (b) Trusting to fate has never turned out as well for me as making a decision to follow a definite course of action.	5
c. Factor III: System Modifiability--measures beliefs about control over racial discrimination, war and world affairs. Item example: 21. (a) One of the major reasons why we have wars is because people don't take enough interest in politics. (b) There will always be wars, no matter how hard people try to prevent them.	5

# of items

d. Factor IV: Race ideology--Items phrased in race-related terms and factored as a separate group: These include some items included in other factors. 13

Item example:

7. (a) Many Negroes who don't do well in life do have good training, but the opportunities just always go to whites. (b) Negroes may not have the same opportunities as whites, but many Negroes haven't prepared themselves enough to make use of the opportunities that come their way. 7

2. Level Appropriateness: The scale was designed for use with adults (as opposed to children) and has been administered to college students as well as high school students.

3. Sample characteristics: There is no indication that standardization was done. The sample on which a factor analysis was based contained 849 males and 846 females, half of a random sample from 12 predominately Negro colleges in the United States.

4. Floor and Ceiling Effects: There is no information on the distribution of test scores.

5. Test-Retest Reliability: No information was reported.

6. Validity data: No information was reported.

7. Administration: The test can be group administered by a teacher or proctor with no special training being required. Test can be completed in 15-20 minutes and can be machine - or hand-scored.

Name of Test: Rotter's Internal-External Locus of Control Scale

Location: Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement.

Test Level: High school and college students

1. Test Content: The scale consists of 23 question pairs, using a forced-choice format, plus six filler questions. Internal statements are paired with external statements. One point is given for each external statement selected. Scores can range from zero (most internal) to 23 (most external).

   Item examples: 7. (a) No matter how hard you try some people just don't like you. (b) People who can't get others to like them don't understand how to get along with others.

   21. (a) In the long run the bad things that happen to us are balanced by the good ones. (b) Most misfortunes are the result of lack of ability, ignorance, laziness, or all three.

2. Level Appropriateness: The scale has been most frequently used with college students, but has been used with adolescent and older subjects. No upper or lower age limits have been established.

3. Sample Characteristics: The scale has been administered to numerous samples, including samples with minority students. However, no evidence could be found that this scale has been standardized.

4. Floor and Ceiling Effects: There is no information on the distribution of test scores.

5. Test-retest Reliability: A reliability coefficient of .72 was computed after one month; after two months r=.55. The one-month sample (N=60) was administered the scale in a group. The two-month sample (N=1;7) completed the scale individually.

6. Validity Data:

   a. Content validity: No information was available.

   b. Concurrent validity: No information was available.

7. Administration: The scale is self-administered and can be completed in about 15 minutes. No special training is required by proctor supervising test administration, and tests can be hand - or machine-scored.

Name of Test: National Longitudinal Study Locus of Control Items

Location: Conger, A.J.; Peng, S.S.; and Dunteman, G.H. National Longitudinal Study of the High School Class of 1972. Group Profiles on Self-Esteem, Locus of Control, and Life Goals.

Test Level: High school and postsecondary

1. Test Content: Locus of control is measured by four items administered as part of a larger questionnaire. Each item has five response options ranging from "disagree strongly" to "agree strongly". Items were scored so that disagreement, indicating an internal locus of control, was scored.

   Item Example:

   1. Good luck is more important then hard work for success.

2. Level Appropriateness: The NLS items were administered to high school seniors and again to these students one year later.

3. Sample Characteristics: The sample consisted of over 15,000 high school seniors of different ethnic/racial background and of different socioeconomic status.

4. Ceiling and Floor Effects: Means and standard deviations are presented in the table below.

Subjects	N	Mean	S.D.
Males	7598	3.68	.73
Females	7933	3.83	.71
Blacks	1892	3.48	.82
Whites	11907	3.81	.69
Hispanics	673	3.49	.80

5. Test-Retest Reliability: The only reliability data presented was for internal consistency.

6. Validity data: No data are reported.

7. Administration: The items are self administered in group format along with the rest of the NLS questions. Administration time is very brief. Results can be hand- or machine-scored.

Name of Test: Motivational Style Scale (MSS)

Author: Dr. Dean R. Spitzer, Department of Education, University of North Carolina at Wilmington

Test Level: College

1. Test Content

The instrument contains 80 items that require the student to "agree" or "disagree" on a 5-point scale (e.g., strongly agree, agree, not sure, disagree, strongly disagree). In using the instrument with incoming freshmen at the University of North Carolina at Wilmington, 3 subscales were identified, and they are: Need for Achievement, Social Need, and Locus of Control. Examples of subscale items are as follows.

a. Need for Achievement:

8. I prefer non-competitive activities.
27. I think that grades should be awarded on the basis of effort rather than achievement.
35. Most of my needs in life are already satisfied.

b. Social Need:

9. Love and companionship are the most important things in life.
32. I love being appreciated by others.

c. Locus of Control:

66. I wish that the world was not so complicated.
80. I feel increasingly helpless in the face of what is happening in the world.

2. Level Appropriateness

The instrument is designed for use at the college level. However, with slight modifications, the instrument could be used at the high school level.

3. Floor and Ceiling Effects

The means and standard deviations of the subscales suggest that there are no floor or ceiling effects.

4. Reliability Data

a. Test-retest Reliability: Not available.

b. Internal Consistency Reliability: see table below.

	Alpha Coefficient
Locus of Control Scale	.75
Social Need Scale	.67
Need for Achievement Scale	.73

c. Alternate Form Reliability: Not applicable.

d. Inter-Level Reliability: Not available.

5. Validity Data

a. Content Validity: There is no information on this criterion.

b. Concurrent Validity: There is no information available on this criterion.

6. Sample Characteristics

The instrument was tested on 524 incoming freshmen at the University of North Carolina at Wilmington. The author (Dr. Spitzer) characterized these students as "mostly from Southeastern North Carolina." A large proportion of the tested students had "serious learning difficulties," according to the author. Only a small proportion of the freshmen class was composed of black students.

7. Administration

The MSS can be administered to a group of students in 15 to 20 minutes. The instrument can be administered by a teacher or proctor and can be scored manually.